5-2019

A Corpus Linguistic Analysis of YouTube Coming Out Videos

Mikhail Zaikovskii
St Cloud State University

A Corpus Linguistic Analysis of *YouTube* Coming Out Videos

by

Mikhail Zaikovskii

A Thesis

Submitted to the Graduate Faculty of

St. Cloud State University

in Partial Fulfillment of the Requirements

for the Degree of

Master of Arts

in English: Teaching English as a Second Language

October, 2019

Thesis Committee:
Edward Sadrai, Chairperson
Choonkyong Kim
Maria Mikolchak

Abstract

With the spread of the Internet and social media, researchers were presented with a novel object for investigation – *YouTube* Coming Out videos. Scholars quickly took up on scrutinizing the phenomenon from various perspectives: as a rhetorical action, an online sanctuary, a tool for developing and spreading the gay collective consciousness, etc. However, in the evolving diversity of studies on *YouTube* Coming Out videos, I failed to find any that are concerned with corpus linguistic analysis, which is highly instrumental in disclosing linguistic trends and unusual characteristics of the texts. Therefore, the main aim of the current study is through the means of corpus linguistics to investigate specific lexemes and collocations that have been used by YouTubers in their Coming Out Videos. More specifically, the study focuses on discovering the distribution of lexical items and collocations in the speech of the YouTubers and pinpointing major thematic groups that emerge from these keywords as a result of general qualitative coding. For the purposes of the current study, two hundred and four coming out stories were selected, vetted, and transcribed into the machine-readable format. The transcripts were further analyzed by the medium of corpus linguistics software that enabled revealing lists of keywords, frequencies, collocations, and concordance lines. Redistributing the most frequently occurring single- and multi-word keywords led to identification of emergent properties – in my case, major themes discussed by the narrators. Among the themes this study identified are *Family, Education, Relationship, Social Media, Vlogging, General Gay-Related Items, Sexuality, Coming Out, Profanity, Homophobia, and Religion*. The pinpointing and analysis of the themes and frequent collocations have expanded current studies on *YouTube* coming out narratives and facilitated better understanding of the contents and rationale behind sharing such deeply personal stories.

Keywords: gay language, *YouTube* Coming Out videos, coming out, corpus linguistics, language and sexuality.

Acknowledgments

To Dr. James Robinson, who believed in me enough to give me a chance to prove that I belong in his program. I cannot even fathom how much I am indebted to him.

To Sofia Logvinenko, who introduced me to Saint Cloud State University. I still remember our very first conversation about the school on a sunny summer day in Moscow. She singlehandedly changed my life on that day.

Table of Contents

Page

List of Tables 6

List of Figures 7

Chapter

I: Introduction 8

II: Literature Review 11

Language and Gender 11

Queer Language 15

The Phenomenon of Digital Coming Out Narratives 18

Corpus Linguistics 22

III: Methodology 27

Searching for the Videos 27

Forming the Pools 28

Analysis 29

IV: Research Findings 32

Description of the Corpus 32

Frequency 34

Keywords 37

V: Analysis of Collocations and Concordances. The Discussions of The Analysis 48

Mutual Information 48

Gay 49

Chapter Page

Year and Grade 56

Mom and Dad 63

God and Religious 71

In Summary 79

VI: Limitations and Further Research 87

VII: Conclusion 90

References 94

List of Tables

Table Page

1. Most frequent lexical words and lemmas 35
2. Single-word Keywords 39
3. Single-word Keywords Themes 41
4. Multi-word Keywords 42
5. Multi-word Keywords Themes 44
6. Single-word vs. Multi-word Thematic Groups 45
7. Collocates of Gay 49
8. Collocates of Year 56
9. Collocates of Grade 58
10. Collocates of Mom 63
11. Collocates of Dad 65
12. Collocates of God 72
13. Collocates of Religious 74

List of Figures

Figure Page

1. Collocates of Gay 51

2. Collocates of Year 57

3. Collocates of Grade 59

4. Collocates of Mom 65

5. Collocates of Dad 67

6. Collocates of God 74

7. Collocates of Religious 76

8. Collocational Network 89

Chapter I: Introduction

Carrying out an empirical investigation in the field of sociolinguistics encompasses a great deal of intertwined theories of interdisciplinary history and research. When it comes to language and gender, or language and sexuality, it is vital to go back to the historical background of the relationship between the terms. Researchers and thinkers have been pointing out the disparities between men and women for ages, so the aptly phrased book title *Men Are from Mars, Women Are from Venus* (Gray, 1992) has today become a ubiquitous and notorious metaphor implying that the two sexes are so dissimilar that they must have come from different planets.

Regarding gender differences, centuries of oppression of women by men left an imprint on the ways both sexes use the same language. These changes have been labeled and conceptualized differently by different authors. For instance, Jespersen (1922) and Lakoff (1973) described female language from the standpoint of a deficit, meaning that there is a standard language that belongs to men, whereas women's language is lacking compared to the standard form. Tannen (1990), on the other hand, distinguishes males and females as representatives of different subcultures – the notion that conveys a diversity of communicative styles between sexes. Such an interpretation contributes to the 'difference' approach towards gender and language.

Furthermore, there are social strata (i.e., LGBT people) the oppression towards which has been, and in some countries is, institutionalized for centuries. Even in the United State, the last anti-sodomy laws were repealed only in 2015. Undoubtedly, the status quo that remained unquestioned during such a long period of time has left behind a print on the evolution of the relationship between language and sexuality (Kulick, 2000). Furthermore, overall, sociolinguistic

study of the language behavior of queer people was seriously hampered due to the public fear of being associated with the outcast community and negative associations the community happened to convey. More specifically, Allen Read (one of the first linguists dealing with language of homosexuals) had to publish his research on texts from public restroom walls in France because he worried that the more conservative American public would perceive his endeavors as pornography (2000).

Consequently, Queer Theory and Queer Linguistics, as independent fields of study, started to emerge only after the gay and lesbian rights movements in the USA began to consolidate in the 1970s and 1980s. Following Motschenbacher and Stegu (2013), for the purposes of this study, I use the term "queer" without reference to a specific social stratum, but rather as a means to show certain concepts through the lens of sexually marginalized people, or more generally, from a non-heteronormative point of view.

With the proliferation and spread of social media, Queer Theory obtained a fresh phenomenon for inquiry – Coming Out *YouTube* videos. The phenomenon has been scrutinized from various perspectives. Losh and Alexander (2010) see the vlogs as a rhetorical action, whereas Wuest (2014) describes the stories as an online sanctuary for queer youth that have been abandoned by major social institutions, e.g. family, school, and church, to name a few. However, in the growing diversity of research on Coming Out *YouTube* videos, I could not find any that deal with corpus linguistics. In the light of the foregoing, the main aim of this study is with the aid of corpus linguistic analysis to examine specific lexemes and collocations that have been employed by YouTubers in their Coming Out Videos.

With this end in view, I selected two hundred and four *YouTube* Coming Out videos posted within the period from 2008 to 2019. Further, the videos were transcribed and analyzed through corpus linguistics software. The computer programs enabled pinpointing keywords, which, in turn, were rearranged into major thematic groups. An exhaustive examination of frequently occurring thematic single- and multi-word keywords by themselves or within the strongest collocations has been used to shed light on the content of the videos. The findings of the current study shall become instrumental in further scrutinizing the phenomenon of the digital coming out narratives through the lenses of Sociolinguistics and Queer Theory as well as potential employing for the benefits of mental health counseling and overall better understanding the mechanism of the coming out.

Chapter II: Literature Review

This chapter delineates historical and theoretical contexts of the current study. The literature is organized around the main aspects which are instrumental for evaluation of the matter of language and sexuality through the lens of corpus linguistic analysis: language and gender, queer language, the phenomenon of digital coming out narratives, and corpus linguistics. More precisely, subsections on language and gender and queer language provide perspectives of various researchers who were among pioneers in creating a conversation about language dissimilarities on the basis of sex. The phenomenon of digital coming out narratives part deals with the reasons behind the emergence and proliferation of deeply personal stories that have millions of viewers on *YouTube*. And finally, subsection regarding corpus linguistics outlines the path of a relatively new field of linguistic analysis that was initially pinpointed and defined in the 1980's and along with the skyrocketed advances in information technologies found abundant applications in various fields.

Language and Gender

This study is concerned with aspects of language specific to sexuality. To my knowledge, the complex interplay between language and sexuality has not been explored on its own terms, but rather as an evolutionary component of the research on the subject of language and gender. One of the first studies that discusses the distinction between speaking styles based on the two conventional sexes was conducted at the beginning of the 20th century by Otto Jespersen, a Danish linguist who specialized in the grammar of the English language.

In his book *Language: Its Nature and Development, and Origin*, Jespersen (1922) created a context which implies that female's language is deficient relative to a male norm. The term

'deficient' was frequently employed in regard to the passage of Jespersen's authorship, among other authors, by Johnson (1983) and Cameron (1998). Jespersen begins the chapter called "The Woman" with a reference to a few historical ethnolinguistic studies that were carried out in the 17th century. Rochefort, the author of the scientific paper published in 1665, reported on a Caribbean tribe in which men and women used distinctive lexemes and collocations: "the men have a great many expressions peculiar to them, which the women understand but never pronounce themselves. On the other hand, the women have words and phrases which the men never use, or they would be laughed to scorn" (p. 237). It is of importance to note that the differences in the same language come to such a degree that Jespersen concludes: "thus it happens that in their conversations it often seems as if the women had another language than the men" (p. 237).

Furthermore, Jespersen provides a number of differences in the English language spoken by his contemporary men and women. For instance, he describes females being more talkative than males. According to Jespersen (1922), loquacity is an essential humans' trait that plays an important role in first language acquisition: "If men had to attend to their children, they would never use so many words – but in that case the child would scarcely learn to understand and talk as soon as it does when it is cared for by women" (p. 142). Through the mother's and nurses' perpetual talks, a child has no choice but to acquire all the necessary linguistic structures given that "the language comes to him as a fresh, ever-bubbling spring" (p. 142).

Jespersen (1922) sees women as perfect imitators in a sense that since early childhood, girls use their "general receptivity, … great power of, and pleasure in, imitation, their histrionic talent, if one might say so" (p. 146) to fathom more efficiently other people's way of talking.

One of the possible consequences of such behavior is more obedient manners among females. Everything that is conventional in language as well as ability to be in agreement with people around are the women's strong suits; whereas, men grow up to be independent actors and thinkers (p. 146).

According to Johnson (1983) and Cameron (1998), Jespersen's rhetoric encouraged perception of female speech as deficient compared to male. A men's talk is seen as a 'norm' thus the context in which one gender is inferior in terms of the other was established. Jespersen's idea that females' speech is deficient compared to a men's norm had been neither largely challenged nor supported until Lakoff published her article *Language and woman's place* in 1973 (Wolfram & Schilling-Estes, 2006). The work was expanded to a book form in 1975, and further the original text from the book was published along with commentaries edited in 2004 by Bucholtz (Bucholtz, 2004).

Lakoff, in the words of Bucholtz (2004), by conducting such a groundbreaking study, crafted an original work of fusion between gender studies and linguistics and created thereby subfield of language and gender studies (LGS). Moreover, she laid the foundations for including LGS in such disciplines as anthropology, communication studies, education, psychology, and sociology (Bucholtz, 2004). Lakoff identified and analyzed linguistic and discourse markers of human speech and asserted, as was also argued by Jespersen, that they are diverging on the basis of gender differentiation.

The spectrum of the linguistic traits and discourse markers typical and in many instances unique to women, according to Lakoff and Lakoff (2004), includes nine items.

1. Women own a large collection of lexemes that are highly specific to their lifestyle, e.g. magenta, shirr, dart (for sewing), etc.
2. So-called "empty" adjectives are being used by women more often than by men. Such words convey vague or little meanings, for instance, divine, charming, cute, etc.
3. Raising intonation in declarative phrases and sentences: "What's your name, dear? Mary Smith?"
4. Extensive use of hedges: "well," y'know," "kinda,", "I guess," etc.
5. Women tend to overuse "so" as in "I like him so much."
6. Girls from early age start using hypercorrect grammar. This point echoes Jespersen's findings discussed above in this section.
7. It is obligatory for a lady to be polite. As a consequence, women tend to overuse superpolite forms.
8. Women do not tell jokes as well as they are not supposed to get jokes.
9. Females speak in italics, meaning they put more stress on an important part of an utterance out of fear that the full statements would not be listened to.

As oppose to the "deficiency", Deborah Tannen became a major proponent of the "difference" approach. In Tannen (1990), she argues that men and women belong to two very distinct "subcultures" from early childhood. Boys, for instance, are placed in the environment of competitive games that later results in developing one-upmanship. Girls, on the contrary, are being encouraged to seek equality and intimacy among each other. Growing up in surroundings that significantly differ from each other, boys and girls acquire completely different communication patterns. To elaborate on that, Tannen introduces the terms 'symmetry' and

'asymmetry.' She argues that the males' main aim of any sort of interaction is to establish and preserve the stance of dominance which creates an asymmetric communicative pattern. Females, on the contrary, tend to focus on symmetry, creating points of convergence using statements like "I know how you feel. I've felt that way before." (Tannen, 1990; Columbaro, 1990).

An important outcome of such studies as those done by Lakoff (1973) and Tannen (1990) is that, by establishing the "difference" and especially "deficit" approaches towards human speech, they not just influenced and altered various fields of study but also set up a concept of gendered binary distinctions. Binary distinctions between male and female talks kindled heated debate in academia due to its extreme nature. The bottom line is the approach created by Lakoff seemed too black and white, a place where a vast number of queer people is missing. The abundance of marginalized non-normative speakers that do not fit into the binary framework laid a groundwork for incipience of Queer Theory, and later, Queer Linguistics (Caskey, 2011; Gaudio, 2004).

Queer Language

Before moving on to exploring Queer Theory and Queer Linguistics that came about as academic evolutionary branches of language and gender studies, I will briefly report on the earlier stages of sociolinguistic endeavors in the field of gay language. More specifically, I will focus on the first half of the 20th century, when sodomy was defined as a sexual crime in many states, and research about language of homosexuals rarely, if ever, extended beyond prisons and psychiatric hospitals.

According to Kulick (2000), likely the earliest recorded evidence of lexemes that were used by homosexual men was collected by Allen Walker Read in the summer of 1928. As a

student (later he would become an English professor at Columbia University and gain popularity for solving the O.K. word mystery [Martin, 2002]), during a sight-seeing trip over the Western United States and Canada, he noted down various texts from public restroom walls. Kulick (2000) draws two major conclusions from Read's findings.

First, the study remained undiscovered by the vast linguistics community due to the author's concerns that the general public would see it as "nothing more than pornography" (p. 248). Read had to secretly print 75 copies of the study in Paris putting a warning "Circulation restricted to students of linguistics, folklore, abnormal psychology, and allied branches of social science" on its cover (Kulick, 2000). This severely hindered possible proliferation of the follow-up studies in the field. The second outcome of the study is the complete absence of the word 'gay' in the entire corpus collected by Read. This fact bolsters "the general belief among etymologists that the term did not exist in its popular meaning of 'homosexual' before the 1950s" (p. 248).

The year of 1941 was marked by the publication of the first English-language lexicon of "the language of homosexuality" (Kulick, 2000). The glossary, which appeared in a two-volume medical issue on homosexuality, consisted of 329 words and collocations. Some of them were used exclusively by the gay community; at the same time, a portion spiraled into the general use, e.g. "drag," "straight," and "basket" (2000). It is important to point out that the glossary disappeared in the later editions of the medical text. The studies that follow on until 1972 deal with an insignificant number of vocabulary – from 26 to 233 terms – yet they are important as far as contribution to the field (2000).

In 1972, Rodgers published a book *The queens' vernacular: a gay lexicon* that contains more than twelve thousand entries (Rodgers, 1972). Kulick (2000) is so impressed by the magnitude of the study, that he asserts "all previous attempts to document gay slang look like shopping lists scribbled on the back of a paper bag" (p. 251). Furthermore, Judith Reisman, despite being a conservative denigrator of homosexuality (Radosh, 2004), also acknowledged the significance of Rodgers's study. In one of her interviews (Tech Consultant, 2017), Reisman states:

> The twelve thousand words … had a certain meaning for the homosexual world [that] the straight world didn't understand. It was another language. [It is] very interesting because … when Webster compiled his American dictionary to distinguish us [Americans] from the English, that had twelve thousand new words, as well, [to show] that we were distinct from the English, from Samuel Johnson's dictionary. Webster said: "We're unique people." He had twelve thousand words to show that. So, this dictionary [*The queens' vernacular: a gay lexicon*] says: 'We are unique people, we have twelve thousand words to show that.' (2017)

The works described above along with more frequent studies and vast gay and lesbian rights movements in the USA in the 1970s and 1980s acted as catalysts to formation of a distinct field of study that was later named Queer Theory (Motschenbacher & Stegu, 2013). The concept of Queer Theory rests on seeing sex, gender, and sexual identities not as static components of heteronormative binary paradigm (as in Lakoff and Tannen, for instance) but as fluid entities that question the idea of fixed gender and sexual identities and challenge the very basis of unified

identity politics" (Piantato, 2016). The adjective 'Queer' when used in a context of Queer Theory or Queer Linguistics, does not really imply deciding what is Queer, but rather aims "to view certain behaviors in a non-heteronormative light or from the perspective of the sexually marginalized" (Motschenbacher & Stegu, 2013, p. 520).

The critical nature of Queer Theory allows to employ its principles within various humanities, e.g. Queer Anthropology, Queer Psychology, etc. However, "as the formation of sexuality-related discourses and categories is primarily a discursive undertaking, Queer Theory proves to possess a special affinity with linguistics" (Motschenbacher & Stegu, 2013, p. 521). Thus, it is hardly a surprise that Queer Linguistics has finally detached itself from Queer Theory, and recently evolved into a more coherent entity (2013).

Leap (2015) charges Queer Linguistics with additional powers to those of Motschenbacher's and Stegu's. From the critical discourse analysis standpoint, he states, it is crucial to identify and scrutinize declarations that take shape of ideological statements. Such statements convey obviousness that people tend to take for granted and pass unquestioned. In this regard, queer linguistics is concerned with "how 'common sense assumptions' about sexuality come to be accepted as 'obvious… right… [and] true' and how the uncritical acceptance of those messages coincides with conditions of difference, hierarchy, and exclusion" (2015, p. 662).

The Phenomenon of Digital Coming Out Narratives

At the beginning of the 21st century, with the rapid spread of social media, gay people were not hesitant to utilize them to question some 'common sense assumptions' of heteronormativity mentioned in the previous subsection. More specifically, Coming Out – the

practice of publicly "revealing stigmatized sexual desire in a heteronormative cultural context" (Zimman, 2009, p. 54) – has been drawing close attention of researchers concerned with human sexuality for over the past three decades (2009); however, the digital video-sharing platform *YouTube*, bestowed queer people with the new quality of the word 'publicly.' For example, a search for 'site:youtube.com coming out story' on *Google* yields over 1 billion results. From the Queer Theory standpoint, Lovelock (2017) argues that through *YouTube* Coming Out stories, "youth are able to articulate what it feels like to be queer in a straight world, and produce and circulate strategies for negotiating a contemporary cultural context defined by increased visibility of LGB [lesbian, gay, and bisexual] identities, alongside the continued dominance of heteronormativity." However, according to Pullen (2010), one specific event was added on to the increase of Coming Out videos popularity.

In February of 2008, a young male teenager of ambiguous sexual identity, named Lawrence King, 15 years old, was murdered at school by Brandon McInerney, his 14-year-old classmate. In response to the homicide, Ellen DeGeneres on her comedy talk show (Lassner, Glavin, DeGeneres, Paratore, and Connelly, 2003–) said the following:

> Days before [the murder], Larry asked his killer to be his valentine. [pause, studio audience responds with emotional shock "ohhh"]. I don't want to be political, this is not political, I am not a political person, but this is personal to me. A boy has been killed and a number of lives have been ruined. And somewhere along the line the killer (Brandon) got the message that it's so threatening and so awful and horrific that Larry would want to be his valentine that killing Larry seemed to be the right thing to do. And when the message out there is so horrible, that to be

> gay, you can get killed for it, we need to change the message. [pause, audience applause]. (as cited in Pullen, 2010, p. 17)

Pullen (2010) asserts that the murder received very limited coverage from the mainstream media. At the same time, it attracted attention by the *YouTube* platform which was gaining popularity at that time. Gay (however, not exclusively) *YouTube* users, by uploading recorded opinions, tried to fulfil the lack of attention to the event from the mainstream media. Furthermore, posted personal stories that resembled the one of King (concerning being outsiders at school, questioning own sexuality, etc.) allowed to feel "copresence, as a sense of being with others" (p. 19). Thus, the power of being able to contribute to an independent empathetic community took part in launching the process of popularization of *YouTube* Coming Out videos.

Losh and Alexander (2010) use an example of a parody on a Coming Out story to assess such stories as a rhetorical action. They argue that the very existence of the parodying video suggests that the online coming out narratives contain similar rhetorical moves, which makes them a genre of its own. Among such common traits, that were employed in the imitation video, the authors identify "familiar tropes that signal both [a person's] hesitance at going public and his sexual availability to like-minded others" (pp. 37-38). Furthermore, the video was set in a kitchen which falls into the category of domestic spaces typical for the genre – e.g., bedrooms, kitchens, living rooms, and dining rooms. The production of the parody can be described as homemade and self-sponsored. Besides, the person positions his rhetoric accordingly to the conventions of the genre: he states that similar videos helped him to come out to some close people prior to filming the video (p. 40).

According to Wuest (2014), Coming Out videos come to the rescue to queer youth that were abandoned by some social institutions, whereas other institutions simply fail to help. He argues that "schools and churches don't bring us in to talk to teenagers who are being bullied. Many of these kids have homophobic parents who believe that they can prevent their gay children from growing up to be gay… by depriving them of information, resources, and positive role models" (p. 31). In this regard, *YouTube* Coming Out stories, as "support structures that queer youth build for themselves in a community of like-minded peers," (p. 31) help young people to overcome feelings of being lonely and neglected.

Lovelock (2017) evaluates the significance of *YouTube* Coming Out videos as objects of analysis and draws two major conclusions. First of all, such narratives offer highly visible cultural texts that may help to reveal important social patterns. And secondly, the researched stories can be used as means to access the contemporary state of the 'homosexual consciousness.' The term encompasses "culturally-specific ideas about what it means to be homosexual/gay in particular social and historical contexts, particularly as these meanings are produced by and amongst LGB-identifying people themselves" (2017, p. 4). Further in the article, Lovelock (2017) attempts to illustrates how exactly *YouTube* Coming Out videos construct homosexual consciousness. Among the milestones along the way of pinpointing the consciousness are self-reflecting on the issue of living in the predominantly straight world and ongoing negotiating and renegotiating of the current cultural context in the light of increasing visibility of LGBTQ+ community.

Corpus Linguistics

One way of analyzing the meanings that are mentioned by Lovelock (2017) is to study the corpus of videos. As can be viewed in the previous sections of this chapter, there have been studies that scrutinize the phenomenon of *YouTube* Coming Out narratives from various perspectives. At the same time, to my knowledge, none deal with the analysis of lexical items and collocations that are specific to the video blogs. Investigation of lexemes helps to unearth the linguistic component of the meanings retranslated through the videos. In this regard, one of the methods to carry out an inquiry of lexical items is corpus linguistics.

According to Lüdeling and Kytö (2008), there is plethora of evidence available to trace the evolution of corpus-based analysis. From the point of view of the contemporary linguistics, corpus is a number of authentic computer-readable texts gathered or compiled according to a set of criteria of any given research design (Brindle, 2016). The same definition can be employed for the pre-electronic corpora, although instead of benefiting from fast and convenient modern electronic devices, researchers had to make tremendously work-intensive efforts that had to be accomplished with pen and paper. Lüdeling and Kytö (2008), and O'Keeffe and McCarthy (2010) identify the 13th century biblical concordances as one of the first significant instances of corpus-based linguistic research. Other examples of pre-electronic corpora are grammars, dictionaries, The Survey of English Usage (SEU) Corpus, etc.

With the development of technology in the mid-20th century, linguists gained access to much larger corpora than ever before. The sizes of the biggest contemporary corpora are estimated to be tens of billions of words. A collection of newly emerged corpus-driven means of analysis required classification and a new name or an umbrella term to refer to. And this is when

the term ‘corpus linguistics’ comes into play. The first accounts of the term date back to the 1980 when Jan Aarts uses it first in Dutch (*corpustaalkunde*) and then, three years later, in English in the title of a collection of papers from the Conference on the Use of Computer Corpora in English Language Research: “Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research” (Lüdeling & Kytö, 2008).

Initially, the term ‘corpus linguistics’ came in handy in underscoring the very specific connotations of linguistic research – computer-based and corpus driven. Nowadays, however, there is a wide range of meanings that has been attributed to the term. More specifically, Taylor (2008) synthesized various views on corpus linguistics in attempt to build a fuller picture of what corpus linguistics has evolved to. Among the alternatives that the author mentions in the article, we can find understanding of corpus linguistics as “a *tool*, a *method*, a *methodology*, a *methodological approach*, a *discipline*, a *theory*, a *theoretical approach*, a *paradigm* (theoretical or methodological), or a combination of these” (Taylor, 2008, p. 180).

In the current study, corpus linguistics is employed as part of the methodology as, for instance, in Brindle (2016), Baker (2003), Pak and Paroubek (2010), etc. In this sense, corpus linguistic analysis incorporates two essential phases: creating the corpora and analysis thereof. Cheng (2011) asserts that most corpus linguists employ only the texts that have been produced authentically. In other terms, researchers see no merit in studying the utterances gathered under controlled conditions. Furthermore, the analysis of corpora is carried out by the means of computer software specifically developed for the purposes of corpus linguistics – e.g., WordSmith, AntConc, Sketch Engine, #LancsBox, etc. It should be noted that this sort of

linguistic analysis deals exclusively with written texts, thus all the other forms of authentic language, such as voice recordings, have to be transcribed prior to studying them.

With the development of corpus software, researchers have acquired access to a multitude of the traits of corpora to scrutinize and to reflect upon. Brindle (2016), Baker (2003, 2004), and Cheng (2011) identified four major tools of the text analysis: keywords, frequencies, collocations, and concordances. Keywords are lexical items that appear more frequently in one corpus when compared against another; this can help to trace lexical items that are unique for the particular corpus. Furthermore, a frequency analysis allows for identifying a range from the most frequently to least frequently occurring lexemes in the corpus. Collocations help to determine meanings and associations between lexemes, which are otherwise difficult to discover by studying individual words in isolation. Through the analysis of concordances, researchers gather information concerning certain categories and contexts, that are conveyed through the lexemes under consideration examined together with the words to their left and right.

Corpus linguistics can be used on its own merits as well as in conjunctions with other types of linguistic research. For example, Brindle (2016) started his study with assembling corpus off the texts from *Stormfront* – a white supremacist far-right online forum. Then, he analyzed the corpus through yielded frequency and keyword lists as well as isolated collocations and concordance lines. Further in his enquiry, he attempted to examine the ways the *Stormfront* forum users employ language to construct identities of heterosexuality and white masculinity. Brindle (2016) discovered that in-group linguistic behavior allows for pinpointing white supremacists' attitudes towards the assorted out-groups – e.g., gay men and women, and racial

minorities. The findings enabled by corpus linguistics were further incorporated in the methodology of critical discourse analysis.

Baker (2003) asserts that corpus linguistic analysis helps to disclose trends and unusual characteristics of the texts, thus creating a solid basis for the discourse analysis of data. In the article, he studied individual personal dating advertisements posted by gay men in *Gay Time* magazine (formerly *Gay News*) in the period from 1973 to 2000. One of the main purposes of the study was scrutinizing the ways employed by advertisers to "negotiate masculine gay identities, both for themselves and in the sort of person they desire to meet" (p. 245). One of the methods that Baker adopted to answer the research questions was corpus linguistics. More specifically, corpus-based examination of frequencies uncovered major ways of constructing and negotiating homosexual identities through the posts in the magazine. In addition, diachronic analysis of the lexical items from the pools of data, distributed on the principle of advertisement publication time, enabled to trace how the constructing and negotiating strategies were undergoing changes overtime.

Various aspects of corpus linguistics have been employed to investigate a wide range of research inquires. For instance, Pak and Paroubek (2010) augmented more traditional methods of corpus linguistics, such as frequencies, concordances, and other methods, with additional means of statistical analysis to conduct "sentiment analysis and opinion mining" in particularly large corpora, such as collections of tweets. Furthermore, in order to examine the potential correlation between cultural-specific genre moves and ultimate efficacy of the letters for the admission purposes, Upton and Connor (2001) had to identify the cultural differences of professional

application letters submitted by American and international candidates. To achieve the goal, the authors employed analysis of concordance lines together with general qualitative coding.

Chapter III: Methodology

This chapter provides a description of the procedure behind collecting data for the analysis and gives a detailed overview of the research tools used in the study. The research questions that I set for this study are as follows:

1. What is the distribution of lexical items and collocations in the speech of gay men in *YouTube* "Coming Out" stories?
2. What major thematic groups emerge from these keywords?

Searching for the Videos

Two hundred and four *YouTube* Coming Out videos were used for the current study. Two pools of videos, consisting of one hundred and five samples each (the total number of videos got reduced to two hundred and four as a result of the vetting procedure described below in the Forming the Pools subsection of this chapter), were chosen using *Google* search filters. The first pool of vlogs consists of the earliest one hundred and five "Coming Out" videos that were posted starting from 2005 (the earliest possible date – the year when *YouTube* was founded). The second pool is comprised of the last one hundred and five videos that had been posted on and before the day I was collecting data for the research. Initially, this search criterion was selected to ensure that my study would deal with two pools that are divided by the longest period of time. However, with the permission of the thesis committee granted during the early stages of carrying out the study, the diachronic portion of analysis was left out for the future research inquiries. By the time the permission was granted, the selection process had been completed based on the original selection methodology that entailed two pools. In this regard, the united pool that has been used for the current study consists of the vlogs posted between 2005 and the present.

More in-depth description of the searching technique proceeds in the following fashion. At the time of writing the thesis, *YouTube* does not provide search commands to meet the criteria set for my study; thus, I had to use the searching capacities of *Google.com*. In order to set irrelevant websites aside, I typed into the search bar the following request: "site:youtube.com gay coming out story." The "site:youtube.com" parameter gives *Google* instructions to only show results from *YouTube*. Furthermore, for the first pool of vlogs, the searching time-span filter was set "From: 01/01/2005." For the second pool, I looked for the most recently posted samples, which requires checking an appropriate box on *Google*.

Forming the Pools

A vlog would be chosen for the study if it were produced by a gay male (excluding transgender individuals) for the purposes of telling his coming out story. Further, it was ensured that the data set excludes interviews, confessions to parent, and any samples that consist of more than one narrator per video. Note, that I selected only the samples filmed by biological males because transgender men (men transitioned from women) language may deviate from the selected population due to different gender backgrounds. The information about vloggers' gender was found either in the titles or in the content of the videos. Only the samples with traceable gender information were selected for the study.

Considering the scope of this study is narrowed to the speech of gay men in a particular domain of language use (i.e., "Coming Out" videos,) and to ensure consistency in data collection and analysis of the subset of the genderlect in question, only the vlogs that are filmed in American (US) English were selected for further investigation. Thus, I excluded all the other varieties from the pools. For the purpose of interrater reliability, two native speakers of

American English, students of the Saint Cloud State University Teaching English as a Second Language program, were consulted. I created a rater checklist that incorporates a list of videos, and four possible options next to each video, considering that American and Canadian varieties can mask each other:

1. Most likely American English
2. Either American or Canadian variety
3. Most likely Canadian English
4. None of the above

If both raters selected options 'three' and/or 'four' for a particular sample, this sample was eliminated from the pool. Only those videos that had been rated as "most likely American English" by both raters were selected for the further analysis.

Considering the nature of the first research question, I am interested in the language that the vloggers use at the time when they speak. In this regard, unless otherwise mentioned specifically, I assumed that videos were posted shortly after they had been filmed. Links for the final selection of two hundred samples were bookmarked in a special Word document and kept on my personal computer.

Analysis

Once the pool of samples was fully formed, I transcribed them using the transcribing tool capacities of *YouTube*. Transcripts were analyzed by means of *Sketch Engine* and *#LancsBox* – software corpus analysis toolkits. The programs were employed to create, for instance, lists of frequencies or collocations, that, in turn, were rearranged in multiple ways tailored to the researcher's specific aims; they also allow for a direct link to the full corpus. In addition,

programs graphically highlight where a particular item (lexeme, collocation, etc.) appears in a text.

More particularly, I used *Sketch Engine* and *#LancsBox* to discover keywords, frequencies, collocations, and concordances. Keywords are lexical items that appear more frequently in one corpus when compared with another, usually bigger one, which can help to trace important linguistic traits of the corpus under investigation. Furthermore, a frequency analysis allows for identifying the most frequently to least frequently occurring lexemes in the corpus. Since language use is not arbitrary, frequency and keyword analyses could potentially reveal emergent categories and trends. Collocations help to determine meanings and associations between lexemes, which are otherwise difficult to discover by studying individual words in isolation. Through the analysis of concordances, linguists examine not mere single lexemes, but rather, certain categories and contexts that are conveyed through these lexemes together with the words to their left and right.

Moreover, I organized the keywords and collocations into specific thematic groups using methods of emergent properties. For this, I employed the methodology of general qualitative coding that enables identifying major themes within the keywords lists. Coding, as described, for instance, in Charmaz (2008), incorporates two major phases: initial (open) coding and focused coding. For the purpose of the initial coding, I went through each item in the lists and labeled the items with a theme they may possibly fit into. More specifically, for the words such as *mom*, *dad*, and *stepdad*, the most suitable theme is Family; whereas, *god* and *religious family* fall into the category of Religion. The items that did not fit in any of the major established categories were put together in the Other category. In order to confirm that the items semantically function

in the manner I attributed to them, I consulted concordance lines. For example, concordance lines **i** and **ii** provide the context based on which I made a decision to code the node **disown** as related to the Family theme.

i. When they're fourteen or fifteen or sixteen years old, they [come out], and their families **disowned** them, and they up on the streets, and all of that.

ii. Dear Mom and Dad, it's been exactly five years since you **disowned** me. Five years of nothing: no contact, no acknowledgement that I even still exist…

After the initial codes had been established, I proceeded to the focused or selective coding. At this stage, major themes were sorted through synthesizing the initial codes. At all the stages of the analysis of the emergent properties, I carried out additional concordance lines to further clarify the semantic belonging of any given item.

Chapter IV: Research Findings

This chapter summarizes the results of the study. In order to layout the findings, I begin with the description of the corpus. The subsequent sections of the chapter explain how by employing two different software packages for the analysis of linguistic data and corpora, *Sketch Engine* and *#LancsBox*, I transformed raw data into the lists of the most frequently used words as well as different types of keyword items. Frequencies and keywords are crucial findings to answer the first research question. Within the unit on keywords, the lists of single- and multi-word keywords were rearranged into major thematic groups in order to find an answer for the second research question.

Description of the Corpus

The corpus is comprised of the transcribed utterances of the *YouTube* vloggers who posted their Coming Out videos from the period of August 11th, 2008 to March 4th, 2019. The total number of videos is 204. The size of the corpus is 384,799 tokens – single occurrences of a word form in the corpus; term 'token' usually refers to 'word'. The number of tokens is represented by 9088 lemmas – all inflected forms belonging to one stem.

The smallest single video incorporates 179 tokens, the largest – 4076 tokens. The mean average count of tokens per video is 1886. Below is an example of raw data (excerpt); a piece has been extracted from the video posted on June 14th, 2011:

> What's up YouTube! So, this is my first video ever! I know it's crazy, but all for this special occasion, I figured that I would do one on coming out. And the reason for this is that when I was younger, when I was about fifteen years old, I would come to YouTube and try to find advice about coming out of the closet. Like, who

> to tell, how to do it, when to do it. And I found that. I got a lot of good advice out of this. And I figured that somehow I could try and bestow some of that knowledge on you guys. As well, you might be a lot younger than me or maybe unfortunately, if you haven't come to terms with your sexuality, you might be a lot older, but, hopefully, I can help you in some way because that… that would really be awesome and something that I'm very interested in. And so, here it goes, I guess.
>
> I first realized that I was gay when I was about fifteen years old. I went to snowboard camp and um… It was so much fun! But there was this guy there, and I got a crush on him. And it was… it was really weird. I really didn't know what it was like. Before that… it would… I had a girlfriend, and we dated for a few months. But um… so that was, it was really strange to like somebody of the same gender as me. Um, so, I come back for snowboard camp and I figured that it was just a phase, like no big deal, but, um, I guess that was when my sexuality was beginning to develop, or that was when I was realizing that I was gay.

At the beginning of the excerpt, the narrator lays out his rationale behind creating the video. For the reason we do not know, he decided to seek help on the matter of his sexuality not from his parents or fiends, but from *YouTube* videos that had already been posted previously: "when I was about fifteen years old, I would come to *YouTube* and try to find advice about coming out of the closet." According to Wuest (2014), more traditional social institutions, such as family, school or church oftentimes fail to create effective support system for young LGBTQ people. Thus, such support structures had to be formed by the queer youth themselves, among

other things, through the means of *YouTube*: "And I figured that somehow I could try and bestow some of that knowledge on you guys." Then, the narrator shares personal story of discovering his sexuality.

This short excerpt that consists of 297 tokens provides the synoptic view of the texts constituting the corpus. Further, I describe how the tools of corpus linguistics have been used to describe the distribution of lexical items and pinpoint major thematic groups of the collected narratives.

Frequency

Through frequency word lists, corpus linguists identify linguistic traits that occur repeatedly in corpora. Usually, a frequency report consists of a word list with the measures of occurrences processed separately for each lexical item. Frequency is one of the most complex and significant concepts in corpus linguistics. Frequency enables yielding and observing specific data representations of corpora; it also serves as a foundation for basic instruments of corpus linguistics – e.g., keywords and concordances. Language is viewed as not an arbitrary concept, thus analysis of frequencies can help to shed light on the patterns of language use. Also, the choice of words and phrases is not neutral, thereby communicating people's cultures and ideologies. Furthermore, frequency is an important indicator of both explicit and implicit lexical preferences. That is to say, researchers use data to draw conclusions from the most frequently occurring lexemes (explicit preferences) in the corpora, as well as from the least frequently occurring or missing lexemes (implicit inferences).

Due to a considerably wide range of lexical items in the corpora, corpus linguists almost inevitably have to deal with an important dilemma: what items "deserve" to be selected for the

further analysis and what lexemes are to be filtered out. Since I intend to use corpus linguistics, among other things, to reveal major thematic groups that have been brought up in Coming Out videos, I decided to look for the most frequent 'lexical words.' According to Brindle (2016), lexical words include nouns, lexical verbs, adjectives, and lexical adverbs.

Within this framework, I compiled two distinct lists of frequencies. The first one incorporates the thirty most frequent lexical words in whatever declension they appear in the text. The second list includes the thirty most frequently occurring lexical word lemmas. Lemma is a canonical or dictionary form of the same lexeme. For instance, "girl," "girls," "girl's," and "girls'" are forms of the same lexeme, with "GIRL" being the lemma.

Table 1

Most frequent lexical words and lemmas

Rank	**Word**	**Frequency**	**Rank**	**Lemma**	**Frequency**
1	I	26912	1	I	26912
2	know	4557	2	KNOW	5434
3	gay	2080	3	OUT*	2384
4	people	1879	4	GO	2348
5	want	1218	5	COME	2334
6	time	1127	6	TELL	2141
7	kind	1114	7	GAY	2112
8	think	1079	8	SAY	2043
9	guys	993	9	PEOPLE	1905
10	tell	951	10	THINK	1759
11	school	893	11	WANT	1709
12	said	867	12	GUY	1414
13	go	853	13	TIME	1249
14	friends	849	14	FRIEND	1247
15	come	843	15	TALK	1128
16	mom	821	16	THING	1031
17	say	799	17	MAKE	1024

Table 1 (continued)

18	video	728	18	VIDEO	924
19	life	713	19	SCHOOL	916
20	love	675	20	MOM	863
21	family	571	21	LOVE	848
22	started	561	22	START	843
23	talk	560	23	SEE	813
24	thing	525	24	YEAR	784
25	day	522	25	LIFE	745
26	parents	517	26	DAY	616
27	story	497	27	FAMILY	603
28	dad	484	28	GIRL	587
29	person	484	29	LOOK	568
30	make	483	30	PARENT	544

* Even though "OUT" is not a lexical word, I deliberately included it into the table for two reasons. First of all, its frequency is exceptionally high. In addition, the word represents one of the key lexemes of the current project about Coming Out videos.

Table 1 demonstrates the findings of the frequency analysis of the corpus. More specifically, we can see that the word and lemma 'gay' are located very high in both lists. Furthermore, the noun lemma that generally represents males constitutes considerably larger portion of the sample than the lemma for females: GUY (with 1414 occurrences) as oppose to GIRL (587). Moreover, lemma MOM (863) is presented in the selection whereas DAD is not. At the same time, the word *dad* scores considerably (almost twice) lower than *mom*: 484 and 821 occurrences, respectively. As for implicit inferences, it is worth pointing out that the lemma WOMAN in not presented in the sample at all. Among the words that convey feelings, *love* and LOVE are the only representatives in both tables. Additionally, *god* and GOD, despite being lower in the lists, constitute one of the major thematic groups of the corpus (see the following subsection).

Keywords

Keywords analysis is based on comparing two corpora. Typical types of corpora to compare are smaller and larger in size, or newer and older. That is to say, lexical items that are used more frequently in one corpus when juxtaposed against another are referred to as keywords. Such items are of particular interest for the study of corpus linguistics because they provide additional merit to investigation. Namely, keywords convey a larger degree of significance in the studied corpora, thus being worth closer scrutiny as oppose to mere frequency analysis. For instance, if an item has higher frequency ranking in a corpus under investigation, it can be due to the item's higher occurrence in the language in general. Therefore, the significance of the item for this particular corpus might be obscured. Keywords, on the other hand, represent a frequency "anomaly" that is unique particularly to the corpus under consideration. *Sketch Engine* allows for the automatic computation of keyness ranking. For this study, I employed 'single-words' and 'multi-words' types of keywords elicitation. The former type shows individual tokens that appear more frequently in the focus corpus than in the reference corpus. The latter type identifies phrases which likewise occur more frequently in the focus corpus when compared against the reference one, but in addition, extracted expressions are listed only if they form collocations that appear to be typical for the language. The reference corpus used is English Web corpus 2015 (enTenTen15). EnTenTen15 was created based on the texts collected from the Internet in the year of 2015. It consists of fifteen billion words where spam and advertisements have been removed from the selection.

To the best of my knowledge, corpus linguistics does not imply the gold standard or the suggested limit for the number of studied items. Brindle (2016) set the cut off point for the

keywords table at 40 because, according to him, "beyond this point the keywords which appeared were semantically similar, [and]… no further semantic groups of words were found" (p. 63). *Sketch Engine* default settings allow for one thousand key single-words and one thousand key multi-words. For the purposes of the current study, I decided to set a limit of one hundred items for both key-words and key-phrases. The matter is that similarly to Brindle (2016), I found that further increasing the numbers of studied lexical items would have brought more repetitions in the themes that have already been established by the first one hundred items.

Table 2 below represents the first one hundred single-word keywords. The head incorporates Frq (Frequency) – the total number of times a token occurs in the studied corpus; Ref Frq (Reference Frequency) – the total for the token's appearances in the EnTenTen15. Furthermore, Score (or keyness) is a feature calculated using simple math, which is a method that allows identifying the range between the most typical lexical items and the rarest ones when one corpus is compared against another. The keyness score of a lexical item is computed according to the following formula:

$$\text{Score} = \frac{fpmFOC + N}{fpmREF + N}$$

where

fpmFOC is the normalized per million frequency of the item in the focus corpus,

fpmREF is the normalized per million frequency of the item in the reference corpus,

N is the smoothing parameter. The default value for *N* in Sketch Engine is 1 (Kilgarriff et al., 2014).

In general, the statistics employed for the keyword analysis by the *Sketch Engine* developers can be simplified to "word W is so-and-so times more frequent in corpus X than corpus Y" (Kilgarriff et al., 2014).

Table 2

Single-word Keywords

Rank	Term	Score	Frq	Ref Frq	Rank	Term	Score	Frq	Ref Frq
1	um	530.74	717	43308	51	like	25.56	10640	18984641
2	cuz	286.03	223	17261	52	wanna	25.09	50	73307
3	gonna	261.81	1039	162819	53	omegle	24.65	10	872
4	yeah	252.12	915	147335	54	awkward	24.58	63	99367
5	gay	172.98	2109	538216	55	mmm	24.19	12	5014
6	okay	96.73	660	293227	56	Instagram	23.9	75	125597
7	closet	79.64	218	106780	57	basically	23.44	267	502425
8	uh	70.08	106	50913	58	feminine	23.06	46	73453
9	texted	69.3	43	10199	59	nigga	22.89	10	2354
10	mom	65.78	862	579984	60	bro	22.87	21	24320
11	blah	65.52	70	30656	61	scare	22.57	125	235229
12	anyways	65.28	104	54610	62	flamboyant	21.88	15	13747
13	dad	61.13	505	358985	63	anybody	21.04	117	236347
14	grandma	56.78	77	43838	64	youtube	20.88	40	69945
15	bisexual	55.37	82	49536	65	whatnot	20.65	12	9021
16	hey	52.77	272	217225	66	homophobe	20.09	9	2975
17	bitch	51.35	88	60197	67	myself	19.86	598	1356793
18	bye	48.28	57	35881	68	hello	19.86	50	97468
19	stepdad	45.77	20	1959	69	mama	19.7	22	33509
20	guy	45.26	1414	1408110	70	whatever	19.44	429	989938
21	disown	44.94	27	9442	71	literally	19.35	178	402489
22	oh	41.8	673	717022	72	myspace	19.27	10	6247
23	homophobic	40.88	35	21136	73	scared	18.94	21	33182
24	whoa	38.62	18	3360	74	makeup	18.77	60	128498
25	homosex-uality	37.99	78	75812	75	uncondi-tionally	18.63	16	21789
26	masculine	37.44	48	40620	76	ashamed	18.4	33	64477

Table 2 (continued)

27	gosh	36.34	33	23568	77	ado	18.28	12	12584
28	bi	35.02	26	16026	78	hurtful	18.1	13	15415
29	youtuber	34.49	14	674	79	guess	18.08	293	722440
30	somebody	33.77	178	222767	80	roommate	17.38	29	58813
31	sexuality	33.51	131	160596	81	anymore	17.24	133	334787
32	boyfriend	33.29	120	146703	82	yo	17.06	16	25489
33	everybody	32.75	270	358435	83	grandpa	16.98	14	20325
34	snapchat	32.57	34	29822	84	girly	16.91	10	9687
35	fuck	32.48	312	420589	85	really	16.88	2277	6141444
36	gotta	31.97	51	54996	86	know	16.58	5434	14939391
37	bla	30.88	16	5855	87	freakin	16.55	9	7536
38	alright	30.66	58	68555	88	stuff	16.52	368	999562
39	stepmom	30.45	13	1700	89	stereotypical	16.16	14	22301
40	dude	30.33	62	75511	90	cry	16.03	195	538001
41	faggot	29.92	15	5110	91	freshman	15.9	80	212434
42	texting	29.91	40	43267	92	kind	15.73	1132	3267658
43	honestly	29.77	132	184645	93	nervous	15.69	96	262017
44	bawl	29.13	14	4180	94	sophomore	15.69	55	142765
45	tinder	28.63	13	2981	95	eighth	15.55	52	135432
46	freak	27.89	70	96818	96	me	15.41	4544	13440829
47	girlfriend	27.75	110	163222	97	god	15.38	144	410225
48	shit	27.6	106	157560	98	straight	15.36	276	802865
49	cousin	25.83	126	204961	99	tell	15.31	2141	6364911
50	weird	25.78	133	217845	100	prom	15.2	17	33855

The first one hundred single-word keywords were further rearranged into twelve thematic groups using general qualitative coding. To ensure the word function in a sentence and categorize its semantic interpretation, I carefully studied concordance lines that comprise the keyword surrounded by the context of the corpus.

Table 3

Single-word Keywords Themes

No.	Theme	Key-Words
1.	Family	mom, dad, grandma, stepdad, disown, stepmom, cousin, mama, unconditionally, grandpa
2.	Relationship	boyfriend, girlfriend, roommate
3.	Social Media	texted, snapchat, texting, tinder, omegle, Instagram, myspace
4.	Vlogging	bye, youtuber, youtube, hello
5.	General Gay-related items	gay, homosexuality, flamboyant
6.	Sexuality	bisexual, masculine, bi, sexuality, feminine, girly, straight
7.	Coming-Out Experience	closet, scare, scared, ashamed, cry, nervous
8.	Education	freshman, sophomore, eighth, prom
9.	Profanity	bitch, fuck, shit
10.	Homophobia	homophobic, faggot, homophobe
11.	Religion	god
12	Other	um, cuz, gonna, yeah, okay, uh, blah, anyways, hey, guy, oh, whoa, gosh, somebody, everybody, gotta, bla, alright, dude, honestly, bawl, freak, weird, like, wanna, awkward, mmm, basically, nigga, bro, anybody, whatnot, myself, whatever, literally, makeup, ado, hurtful, guess, anymore, yo, really, know, freakin, stuff, stereotypical, kind, me, tell

As a result of the keywords rearranging, I came up with twelve thematic categories (Table 3): *Family, Relationship, Social Media, Vlogging, General Gay-Related Items, Sexuality, Coming-Out Experience, Education, Profanity, Homophobia, Religion,* and *Other*. The largest group that conveys semantic load relevant for the study turned out to be the *Family* group consisting of ten items: mom, dad, grandma, stepdad, disown, stepmom, cousin, mama, unconditionally, and grandpa. The smallest group is *Religion* with only one item – god.

Table 4 below represents first one hundred multi-word keywords. In a similar vein to Table 2, Table 4's head includes Rank, Term, Score, Frq (Frequency), and Ref Frq (Reference

Frequency). The ‘multi-words’ mode helps extract phrases that are more common for the focus corpus than for the reference one, when the former compared against the latter. Additionally, the expressions are listed by the software only if they appear to be collocations typical for the language.

Table 4

Multi-word Keywords

Rank	Term	Score	Frq	Ref Frq	Rank	Term	Score	Frq	Ref Frq
1	coming-out story	455.89	184	79	51	friend group	18.05	7	345
2	being gay	87.42	37	1148	52	gymnastics team	17.86	7	0
3	i kind	79.51	34	1376	53	freshman year of high school	17.52	7	911
4	gay person	78.2	34	1710	54	birth mom	17.48	7	939
5	gay community	67.67	40	8865	55	other girl	17.44	8	3610
6	freshman year	64.21	60	24550	56	text message	17.34	21	37963
7	sophomore year	57.08	44	17131	57	last person	17.27	10	9105
8	eighth grade	54.79	45	19449	58	senior year	17.26	22	40874
9	gay guy	52.56	22	1083	59	bla bla	17.19	7	0
10	coming-out experience	50.64	20	18	60	third grade	16.99	15	23000
11	seventh grade	50.25	31	10147	61	same gender	16.87	8	4352
12	blah blah	46.6	25	6503	62	religious family	16.69	7	1858
13	youtube channel	45.84	22	3931	63	making fun	16.64	9	7419
14	gay man	42.09	26	10254	64	little kid	16.47	9	7677
15	first boyfriend	41.49	17	762	65	big secret	16.45	7	2163
16	sixth grade	41.48	32	17268	66	fourth grade	16.12	12	16745
17	last video	39.48	17	1737	67	boy drama	15.89	6	14
18	blah blah blah	35.94	17	3730	68	mom kind of	15.88	6	24
19	gonna talk	34.74	14	549	69	whoa whoa	15.83	6	83

Table 4 (continued)

20	gonna change	34.61	14	609	70	first relationship	15.5	6	462
21	year of high school	34.47	21	9964	71	guy friend	15.49	6	480
22	gonna start	33.92	14	995	72	gonna look	15.43	6	557
23	next video	33.65	15	0	73	real mom	15.35	6	662
24	coming-out process	33.13	13	84	74	boom boom	15.32	6	695
25	first video	32.55	21	11637	75	second person	15.26	8	6758
26	first person	32.39	50	52646	76	whole life	15.21	21	45856
27	middle school	31.31	109	141116	77	gay culture	14.92	6	1202
28	fifth grade	30.58	25	19532	78	conversion therapy	14.91	6	1230
29	gay kid	30.54	12	150	79	hello everyone	14.71	6	1482
30	entire life	30.37	39	40831	80	english class	14.53	8	8004
31	whole situation	27.78	15	6925	81	high school	14.35	304	949554
32	little bit	26.88	187	299804	82	youtube video	14.3	7	5244
33	junior year	26.26	23	22301	83	great feeling	14.23	8	8574
34	certain way	25.5	19	16349	84	tenth grade	13.95	6	2557
35	gay friend	25.12	10	528	85	single time	13.51	8	10016
36	straight guy	24.52	10	988	86	laramie project	13.41	5	5
37	story time	22.72	13	8540	87	black gay man	13.36	5	84
38	car ride	22.11	11	5158	88	word gay	13.33	5	118
39	huge weight	22.04	9	1098	89	super super	13.3	5	163
40	video today	21.89	9	1215	90	facebook post	13.3	6	3601
41	gay bar	20.88	9	2171	91	whole nother	13.25	5	236
42	next question	19.94	13	0	92	i text	13.21	5	280
43	big deal	18.9	34	64677	93	like kind	13.17	5	344
44	first guy	18.79	8	2039	94	whole time	13.11	13	28290
45	elementary school	18.51	46	96067	95	wonderful day	13.1	8	10899
46	scary thing	18.49	8	2356	96	hello everybody	13.1	5	451
47	ninth grade	18.48	10	7302	97	telling everyone	13.08	6	3961

Table 4 (continued)

48	long story	18.44	15	19732	98	telling everybody	13.02	5	553
49	ass bitch	18.23	7	155	99	like everyone	12.89	5	743
50	mom dad	18.12	7	266	100	toxic side	12.82	5	861

Employing the same strategy as with single-word keyword items, I rearranged multi-word items into thematic groups. In order to classify the word function within a sentence as well as its semantic role, I observed the concordance lines. The results of the categorizing are reflected in Table 5.

Table 5

Multi-word Keywords Themes

No.	**Theme**	**Key-Words**
1.	Education	freshman year, sophomore year, eighth grade, seventh grade, sixth grade, year of high school, middle school, fifth grade, junior year, elementary school, ninth grade, gymnastics team, freshman year of high school, senior year, third grade, fourth grade, english class, high school, tenth grade
2.	Coming-Out Experience	coming-out story, coming-out experience, coming-out process, first person, entire life, huge weight, car ride, scary thing, little kid, big secret, second person, whole life, great feeling, whole time, telling everyone, telling everybody
3.	General Gay-Related Items	being gay, gay person, gay community, gay guy, gay man, gay kid, gay bar, same gender, gay culture, word gay, black gay man
4.	Family	mom dad, birth mom, mom kind of, real mom
5.	Religion	religious family, conversion therapy
6.	Vlogging	youtube channel, last video, gonna talk, gonna start, next video, first video, story time, video today, hello everyone, youtube video, wonderful day, hello everybody
7.	Relationship	first boyfriend, gay friend, first guy, friend group, boy drama, first relationship, guy friend
8.	Social Media	facebook post

Table 5 (continued)

9.	Other	i kind, blah blah, blah blah blah, gonna change, whole situation, little bit, certain way, straight guy, next question, big deal, long story, ass bitch, other girl, text message, last person, bla bla, making fun, whoa whoa, gonna look, boom boom, single time, laramie project, super super, whole nother, i text, like kind, like everyone, toxic side

It is worth pointing out that the list of thematic groups composed for the multi-word items differs from that for the single-word ones (Table 6). Multi-word themes proceed as follows: *Education, Coming-Out Experience, General Gay-Related Items, Family, Religion, Vlogging, Relationship,* and *Other*. Whereas, for the single-word items I composed *Family, Relationship, Social Media, Vlogging, General Gay-Related Items, Sexuality, Coming Out Experience, Education, Profanity, Homophobia, Religion,* and *Other.* Thus, it is apparent that three extra thematic groups have been identified for the single-word items: *Homophobia, Sexuality,* and *Profanity.* The range of each multi-word category is also different. As an illustration, the largest group that comprises relevant semantic load is *Education* with nineteen items in it: freshman year, sophomore year, eighth grade, seventh grade, sixth grade, year of high school, middle school, fifth grade, junior year, elementary school, ninth grade, gymnastics team, freshman year of high school, senior year, third grade, fourth grade, English class, high school, and tenth grade. The smallest group is *Social Media* consisting of just one item – facebook post.

Table 6

Single-word vs. Multi-word Thematic Groups

No.	**Single-Word Group (SW)**	**Number of Items in a SW**	**Multi-word Group (MW)**	**Number of Items in a MW**
1.	Education	4	Education	19

Table 6 (continued)

2.	Coming-Out Experience	6	Coming-Out Experience	16
3.	General Gay-Related Items	3	General Gay-Related Items	11
4.	Family	10	Family	4
5.	Religion	1	Religion	2
6.	Vlogging	4	Vlogging	12
7.	Relationship	3	Relationship	7
8.	Social Media	7	Social Media	1
9.	Other	59	Other	28
10.	Homophobia	3		
11.	Sexuality	7		
12.	Profanity	3		

The scope of the present study does not permit for the exploration and examination of all the keywords from the yielded lists. Thus, I made a decision to include in the further analysis a number of items that appear to be the most prominent representative for the yielded semantic groups. *Gay* is not only one of the most frequent lexical items of the entire corpus, but also it is the key concept of the current thesis, since the study is concerned with the language of gay men sharing their coming out experiences on *YouTube.* Moreover, the largest semantically significant thematic group within the multi-word keyword categorization – *Education* – consists of such frequently reoccurring constituents as *year* and *grade*. For instance, *freshman year, sophomore year, eighth grade, seventh grade, year of high school,* etc. *Mom* scores very high in the frequency list and represents the most popular item in the Family thematic group. As I mentioned in the description of frequency lists, *dad* is located significantly lower in the list than *mom* with the ranks 16 (821 occurrences) and 28 (484 occurrences), respectively.

Finally, it is essential to analyze not only the most popular lexemes, keywords or collocates but also lexical items that occur less frequently or even transcend the yielded lists. In this regard, the topic of religion is worth bringing up with such items as *god* and *religious* (as in *religious family*).

In the light of the foregoing, for the subsequent, more in-depth, analysis of collocations and concordances, the following terms have been selected: *gay*; *year* and *grade*; *mom* and *dad*; and *god* and *religious*.

Chapter V: Analysis of Collocations and Concordances. The Discussions of The Analysis

In this chapter, I provide an overview the results of the analysis of collocations and concordance lines in order to gain a better understanding of how the various lexemes, described in the section on the research findings, influence our view of the corpus. To trace the influence, along with the concordances, I discuss possible explanations behind the high frequencies of the nodes selected for the analysis. The final subsection of the chapter embeds my findings within the context of existing research in this field.

Mutual Information

Mutual Information (MI) is a widely used statistical tool to determine the strength of association between a node and its collocates. I followed recent studies on the matter such as Brezina, McEnery, and Wattam (2015), Brindle (2016), and Baker (2016) to come with suitable levels of the MI parameter for my study. MI can be defined as follows: it is a "measure of the information overlap between two random variables" (Bouma, 2009, p. 32). In relation to Corpus Linguistics, MI is used to determine the semantic distance between collocates. Respectively, the higher MI the stronger association within a collocational network.

Brindle (2016) based his examinations on the assertion that MI that equals or above 3.5 allows for creation the principal links between the lexemes to assemble a collocation. Baker (2016), on the other hand, states that for a collocation to appear "psychologically real", in other words, one lexeme to activate an association for another, the MI parameter is supposed to be set at least as high as 6.0.

Gay

I will start my analysis with the strongest collocates of the node **gay** filtered by the highest MI score. As can be observed from Table 7, the collocates ranking from 1 to 20 convey particularly strong semantic bonds with the node **gay**, since the MI score is above 5.5 for all the instances. **Frequency (collocation)** column incorporates the total number of occurrences of a respective collocate of the node. The total number of the collocate occurrences in the entire corpus is reflected in the **Frequency (corpus)** column.

Table 7

Collocates of Gay

Rank	Collocate	MI score	Frequency (collocation)	Frequency (corpus)
1	representation	7.30	6	7
2	fag	7.26	5	6
3	son's	7.26	5	6
4	bars	7.16	7	9
5	openly	7.05	13	18
6	stereotypical	6.72	8	14
7	rights	6.52	5	10
8	lesbian	6.52	23	46
9	transgender	6.39	5	11
10	bar	6.09	10	27
11	community	6.09	50	135
12	being	5.99	271	786
13	assumed	5.94	6	18
14	culture	5.89	9	28
15	bla	5.85	5	16
16	club	5.76	5	17
17	rumors	5.76	5	17
18	trans	5.69	7	25
19	calling	5.64	10	37
20	closeted	5.63	7	26

The graph below (Figure 1) is designed in such a way so three distinct traits or dimensions of the collocates can be observed:

1) **Strength of collocation** – calculated by the association measure (MI) and presented on the graph by the distance (length of line) between the node and the collocates. To better illustrate the semantic force between the node and the collocate, *#LancsBox* employs the 'magnet effect': the closer the collocate is displayed to the node, the stronger the association between the two.
2) The color of the collocate represents its **frequency** in the entire corpus: the more intense the shade, the higher frequency rank of the collocation.
3) **Position of a collocate** relative to the node reflects a respective position of the collocate in the corpus: to the left; to the right; sometimes to the left, sometimes to the right (middle position in the graph).

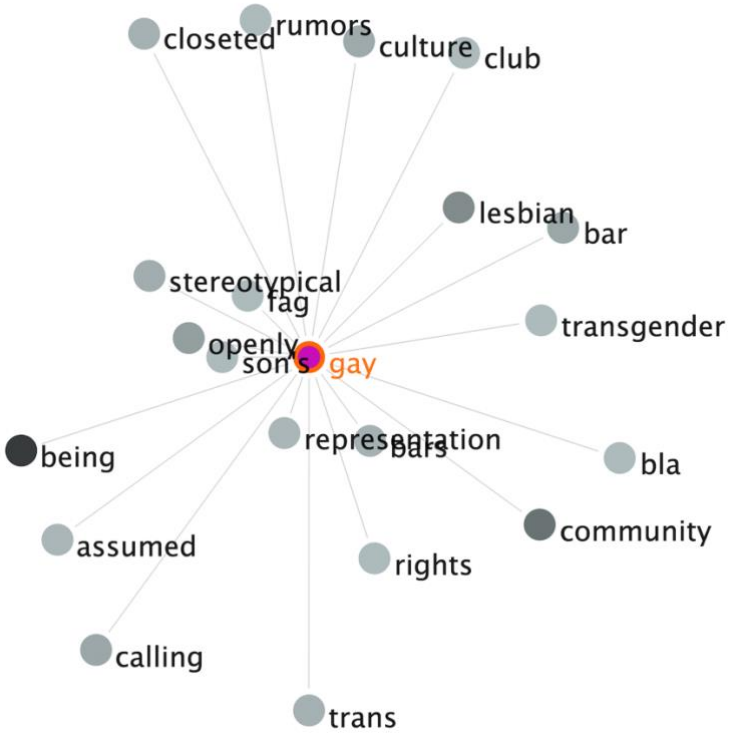

Figure 1. Collocates of Gay

As can be observed in both Table 7 and Figure 1, the most frequent collocation of the corpus is **being gay,** leading by a significant margin: 271 occurrences compared to the second most popular **gay community** with 50 occurrences. **Being gay** is employed to elucidate a wide range of narratives within the topic of homosexuality in the corpus.

From such a high number of occurrences of the collocation **being gay** in the corpus, it can be assumed that the vloggers used the phrase to express a broad spectrum of notions and feelings, as can be examined through concordance lines. For instance, the line one conveys relatively neutral tone – just the expressing a point in time. On the other hand, lines three, four, five, and seven account for negative feelings experienced at different stages of accepting one's sexuality.

1. …getting more details about what it's like for me being in the Navy and **being gay**. I was actually out before "Don't ask don't tell."
2. I really wanted you to know that I'm okay with who I am, and I really hope that you are too, because **being gay** doesn't make me less human, it doesn't make me less worthy of acceptance or of love.
3. I didn't love myself. I hated myself. I hated everything about myself because **being gay,** I just wanted to get over that, and I prayed every night that I would wake up straight.

More precisely, a vlogger used the line two to justify his right to be called a human being. Whereas, the line three is used to share an attitude of denying person's queerness and even hatred towards himself, desire to change the sexuality with help of a prayer or miracle (overnight).

4. I remember going home and crying after school, I was also teased for **being gay**. A lot of people noticed that I was different.
5. We all have our differences and **being gay** is not a choice.
6. It's not an accurate representation of what **being gay** is, I mean for me **being gay** is such a random insignificant detail about the person, that I am **being gay** is the last thing that I identify with.
7. I would pray to God, please, please, don't make me gay! Please, please, make me like everybody else! They would tell me that **being gay** is a sin.

Both line three and line seven incorporate religious-driven hopes to convert to "normal." In the line six, a completely different stage is being described using the same collocation; a man

basically says that his sexual orientation plays the least significant role in the way he identifies himself as a human being ("**being gay** is the last thing that I identify with"). Furthermore, line 8 showcases a positive effect of coming out carried out by the collocation: "**Being gay** has opened my eyes to so many things."

8. Things have gotten so much better since then [coming out]. I feel like I've evolved so much as a person. **Being gay** has opened my eyes to so many things. I feel like I've developed this really great sense of empathy for other people.

From the examples above, we can see that collocation 'being gay' is a particularly important tool employed by gay vloggers to set a stage for expressing attitudes about homosexuality, evaluating concerns and struggles, sharing joy and relief, and so on. The term 'homosexual consciousness' described by Lovelock (2017) encapsulates notions specific to a particular culture about what it means to be homosexual/gay. Narratives produced by LGBTQ+ identifying people constitute the most valuable part of constructing 'homosexual consciousness.' In this regard, it is not a surprise that the present participle 'being' plus 'gay' helps people, having a desire to publicly come out, create a linguistic scaffolding for their own vision of 'homosexual consciousness.'

Gay community is the second most frequent collocation with the node 'gay' in it. YouTubers express assorted feelings and experiences of identifying themselves as a part of gay community. For some of them, affiliation with the group brings pride and happiness, for the others, it is a source of negative associations and experiences. I used five concordance lines to illustrate various instances in which the collocation has been used.

9. I feel, it I was put in its ugliest body possible to shed light to, like, the toxic side of the **gay community**, to shed light on the heart and the shape, on the shaming, and the body shaming…

According to Baker (2003), masculinity has often been perceived by gay men as a particularly desirable feature to obtain. Such desire traces its roots to the historically almost ubiquitous – especially since the onset of Christianity and Islam – negative attitudes towards "womanlike" men. Baker (2003) asserts that the attitudes contributed to the creation of a stereotype, circulating within the general public, which sees homosexuals "as camp of effeminate" (p. 245). Furthermore, a lot of people still cannot come into terms with their own body. They keep feeling insecure even having come out of the closet. For instance, in line nine, a man speaks about the "toxic side of the gay community." He refers to some standards that have been set by his fellow gay men; standards, that at some point in his life he failed to meet.

10. I do wanna thank you very much because I do feel like what I have to say can… even though I'm mainly speaking through gay… the **gay community** and not the LGBT community as a whole; I can only speak for what I know.

Gay community appears a particularly complex entity when discussed by its members. First of all, line ten discriminates gay community from the entire LGBTQ community. Even a few decades ago, there were only gay men and lesbians identified, and all the other shades on the LGBTQ spectrum had to either adhere to the two major categories or try to assert themselves. Whereas nowadays, this vlogger does not feel that he is entitled to speak for the entire nonheteronormative community, meaning that gay people might be to some degree different from other representatives of queer community.

11. It may seem like I'm just making this up because I fit the non-attractive, you know, traits of the **gay community,** but like I said a thousand goddamn times, I feel like I was meant to go through my body struggle on everything blah blah blah.
12. I got too lazy to write down because it's like you emerge from the closet expecting to be this butterfly, and the **gay community**… and the **gay community** just slaps you, just slaps the idealism out of… you.

Line eleven incorporates a viewpoint that **gay community** has some sort of "non-attractive" traits related to a human body. Another negative characteristic of gay community is expressed in the line twelve. The vlogger initially assumed that after all the struggles of being in the closet, the result of coming out would become an easier well-being: you "emerge from the closet expecting to be this butterfly." However, in the reality, the gay community apparently has some cruel forces: "the **gay community** just slaps you, just slaps the idealism out of out."

13. It is something I never thought was possible. I never thought that I would find people like me or that I got along with in the **gay community**, or anything. And like when I first came out, I never imagined that I would be going to drag shows every week or wearing wigs.

For some people, getting along with the gay community is associated with a liberating process. In line thirteen, the man shares his excitement about how around the time of his coming out, he did not anticipate some opportunities that the process of integrating to the gay community may involve. However, now he seems happier: "And like when I first came out, I never imagined

that I would be going to drag shows every week or wearing wigs." He can now afford doing things he did not think he could before coming out.

Year and Grade

The largest semantically substantial thematic category among multi-word keyword classification has to do with education. Vloggers shared a plethora of narratives that contain facts related to specific school or college years: *freshman year, sophomore year, eighth grade, seventh grade, year of high school,* etc. All the instances incorporate two nodes: **year** or **grade**. In this section, I will identify major themes related to the collocations that contain the two nodes.

Table 8

Collocates of Year

Rank	Collocate	MI score	Frequency (collocation)	Frequency (corpus)
1	sophomore	9.98	53	55
2	freshman	9.92	72	78
3	junior	9.73	34	42
4	eighth-grade	9.55	5	7
5	senior	9.47	27	40
6	semester	8.51	8	23
7	college	7.64	26	137
8	last	7.61	27	145
9	beginning	7.36	8	51
10	half	7.27	9	61
11	year	7.16	50	367
12	summer	7.10	11	84
13	throughout	6.95	6	51
14	old	6.91	18	157
15	high	6.81	37	346
16	entire	6.39	8	100
17	during	6.31	6	79
18	ago	6.09	7	108

Table 8 (continued)

19	almost	6.07	5	78
20	until	6.07	11	172

As reflected in Table 8, the five strongest collocates of the node *year* are *freshman* (72 co-occurrences), *sophomore* (53), *junior* (34), *eighth-grade* (5), and *senior* (27). All the collocates are located very close to the node as can be observed in Figure 2. Such a short distance stemming from especially high MI scores (Table 8) that range from 9.47 to 9.98. All the five collocates are notable because they consolidate a cluster of adjectives describing a very specific time of high school or college.

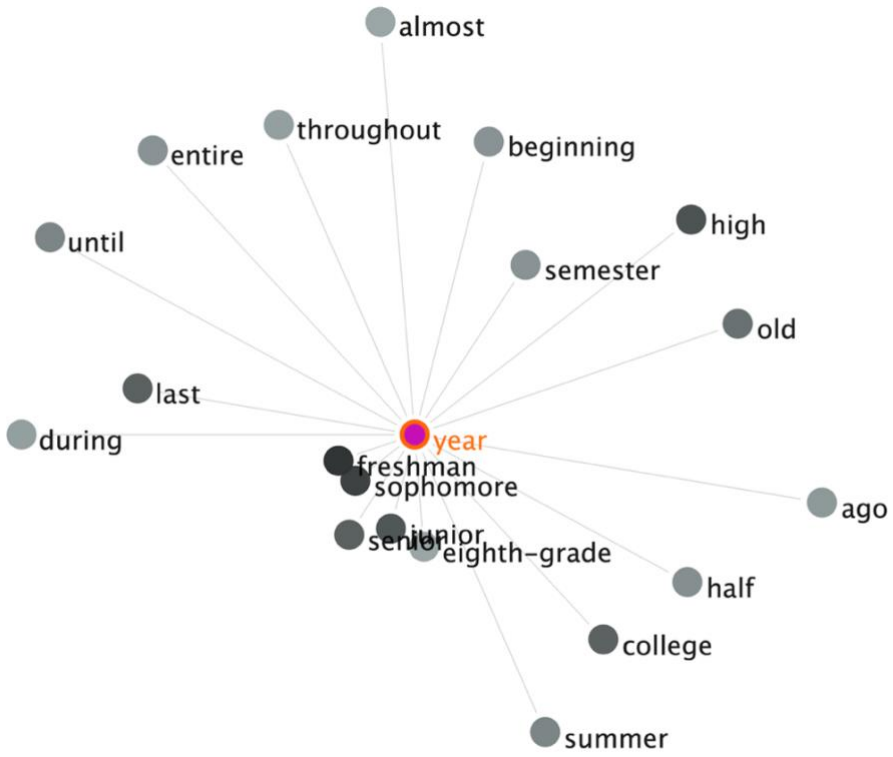

Figure 2. Collocates of Year

The results for the node *grade* are even more consistent in terms of thematic representation. Table 9 predominantly incorporates various numbers that represent years of high school (*ninth, 10th,* and *tenth*), middle school (*six, sixth, 7th, seventh, eighth,* and *8th*), and elementary school (*third, fourth,* and *fifth*). The MI scores for the set of numbers – ranks from 1 to 12 – are particularly high: with 11.35 being the highest, and 9.54 – the lowest. Further, in the concordance analysis, I will try to discover the most significant events that took place in those school years.

Table 9

Collocates of Grade

Rank	Collocate	MI score	Frequency (collocation)	Frequency (corpus)
1	tenth	11.35	9	6
2	seventh	10.92	39	35
3	ninth	10.90	11	10
4	sixth	10.81	36	35
5	8th	10.77	8	8
6	fifth	10.77	28	28
7	10th	10.77	7	7
8	7th	10.77	7	7
9	eighth	10.74	51	52
10	fourth	10.35	21	28
11	third	9.60	16	36
12	7	9.54	6	14
13	grade	7.68	26	221
14	since	7.31	15	164
15	end	7.15	13	159
16	six	7.02	5	67
17	summer	6.96	6	84
18	during	6.78	5	79
19	until	6.66	10	172
20	middle	6.50	7	135

Through the length of lines, connecting the node *grade* and the numbers that represent school years, Figure 4 illustrates the "magnet force" that occurs between the node and the collocates. Shorter lines represent stronger association between the node and the collocate. Furthermore, the more intense shades of circles (for instance, for *eighth* and *seventh*) convey the higher frequency rank of the collocate in the corpus. Table 9 confirms this: the frequency (within the collocation) for *eighth* is 51 and for *seventh* is 39. As opposed to, for instance, significantly lighter shade of *tenth* with the frequency being only 9 occurrences.

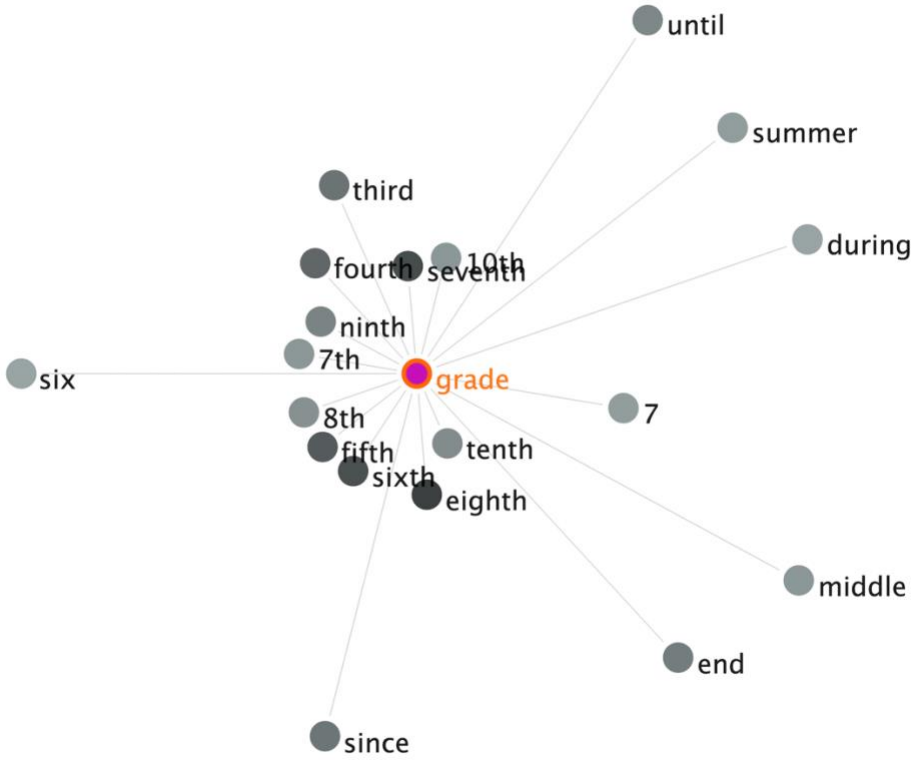

Figure 3. Collocates of Grade

The follow concordance lines of *grade* and *year* are representative of those found within the data.

14. Today I would like to share with you my coming-out story. Um, so it all started like my **freshman year**. I mean, I've always known I was gay. But it… I've just always like pushed it to the back of my head.
15. I really don't know where to start but I guess I'll just tell you my coming-out story. I came out my **freshman year** of high school. I came out to my mom.
16. When I got into college, my **freshman year** that was the first time I actually began to research on sexuality, I did a Google search.

Describing their coming out process, vloggers made a plethora of references about school or college years. It is noteworthy that such references are not made arbitrarily, but rather have a systematic nature. For example, lines 14-16 reveal that during freshman year of college or high school, the men decided either to come out or take a closer look at one's sexuality. In these instances, **freshman year** imbodies not just the beginning of a new academic segment in their lives but also an onset of coming to terms with one's sexuality one way or another. For some young men, freshman year is a point when they start living separately from parents for the first time. This brand-new freedom involves fresh experiences and superfluous responsibilities as well as some important resolution, such as coming out ("I came out my **freshman year** of high school" or "my **freshman year** that was the first time I actually began to research on sexuality").

17. So, um, that brings me to about sophomore year, and that was I think **sophomore year**, mid sophomore year, that was what I finally discovered, I came to terms of my sexuality and I realized that I was gay.
18. So, comes January of my **sophomore year.** This is the first time I really had strong feelings to a guy, and this is openly gay guy.

19. I was **eighth grade**, so, eighth and your **freshman**, and your **sophomore**? I think I finally accepted that I was gay when I was a **sophomore.** Took me a while off.

Obviously, gay men do not stop thinking about their sexuality during sophomore and other years; however, the frequency of the collocate **sophomore** in the corpus is lower than this of **freshman**: 53 and 72, respectively. Furthermore, whereas **freshman** is mostly associated with actual coming out or first romantic or sexual relationship, **sophomore** and **junior** mostly have to do with more general phases of self-acceptance such as awareness ("I realized that I was gay, like there's no question about it" – line 20) or confusion ("even though I knew I was gay, I still liked her so much, but I didn't fully know, I never had a girlfriend" – line 21). At the same time, the collocate **freshman** is not always used to express some sort of a breakthrough, it is also mentioned in describing mundane pieces of information: "I got really bad grades like F's all throughout **freshman year**" (line 22).

20. I guess I first came out to my friends when I was a **junior** in high school. So, I was about 17 years old, 16 or 17, before this. I realized that I was gay, like there's no question about it.
21. … from eighth grade all the way until like piling to your **senior year** of high school, even though I knew I was gay, I still liked her so much, but I didn't fully know, I never had a girlfriend.
22. I did stop smoking because it was stupid, and I didn't see the point of it. So, I got really bad grades like F's all throughout **freshman year**, and ahead of your summer school, and that sucked.

Furthermore, it is not a surprise that the node **grade** and its collocates construct narratives related to the earlier stages of the awareness about one's homosexuality. The point is that the node in essence is being mostly used to illustrate elementary and middle school years. Whereas, **year**, as an educational milestone, has been used mostly in relation to a high school or college. I was not able to find any episodes of coming out linked to the node **grade**. The biggest difference between the contexts of **year** and **grade** is that for the former, the references to sexual attraction and episodes of actual coming outs prevail over those related to realization of one's homosexuality that are more typical for the latter. For instance, "I knew [I was gay] somewhere in between **fourth** and **sixth grade**" (line 23), "I'm pretty sure I first realized it at a pretty young age. I was probably **fourth grade** or something" (line 24), and "I actually began to develop same-sex attraction for one of the guys in my class in the **seventh grade**" (line 25).

23. People think they don't figure it out until they're, like, in high school. Like, that's not me. I knew somewhere in between **fourth** and **sixth grade.** I know, I knew by **sixth grade**, but I think maybe I knew in **fifth grade**, **fourth grade**. I'm not really sure.
24. So, I'm pretty sure I first realized it at a pretty young age. I was probably **fourth grade** or something. I acknowledge that I looked at boys the way, like, boys would talk about girls.
25. I actually began to develop same-sex attraction for one of the guys in my class in the **seventh grade**. Um, I didn't know what homosexuality was.

Mom and Dad

The nodes **mom** and **dad** are located high in the frequency list (seventh and twentieth ranks, respectively) and represent two the most popular items in the single-word key-word Family thematic group. In this section, I attempt to shed light on relationship between vloggers and their parents through the set of strongest collocates associated with the nodes. Tables 10 and 11 each illustrate twenty collocates with the highest MI scores binding the collocates and the nodes **mom** and **dad**, respectively.

Table 10

Collocates of Mom

Rank	Collocate	MI score	Frequency (collocation)	Frequency (corpus)
1	stepdad	7.82	9	19
2	birth	7.80	7	15
3	kitchen	6.96	6	23
4	dad	6.15	70	471
5	texted	6.06	6	43
6	mom's	5.98	7	53
7	grandma	5.92	9	71
8	my	5.86	852	7017
9	cry	5.76	5	44
10	asked	5.49	15	159
11	conversation	5.41	8	90
12	stay	5.38	6	69
13	told	5.29	73	890
14	loved	5.29	9	110
15	crying	5.26	10	125
16	mom	5.25	64	804
17	religious	5.24	5	63
18	she's	5.24	26	328
19	called	5.18	12	158
20	hey	5.12	20	275

As indicated in Table 10, among the first 20 collocates of the node **mom**, the strongest collocate **stepdad** has 7.82 MI score, whereas the last collocate in the list **hey** scores 5.12. Considering that all the MI scores are listed in descending order, each collocate in the table conveys strong association with the node: MI's > 5.00. Figure 4, among other things, provides a visual representation of the "magnetic force" between the node and the collocates. It can be observed in the figure that collocates **stepdad, birth, and kitchen** are located particularly close to the node, which corresponds with the collocates' highest MI rankings: 7.82, 7.80, and 6.96, respectively.

Furthermore, the list appears to reveal that within the corpus a phrase "my mom" (and other combinations of *my* and *mom* that fall into the 5.86 MI threshold) is more than 11 times higher in frequency than the next frequent combination **told** plus **mom**: 852 co-occurrences of **my** plus **mom** and 73 co-occurrences of **told** plus **mom**.

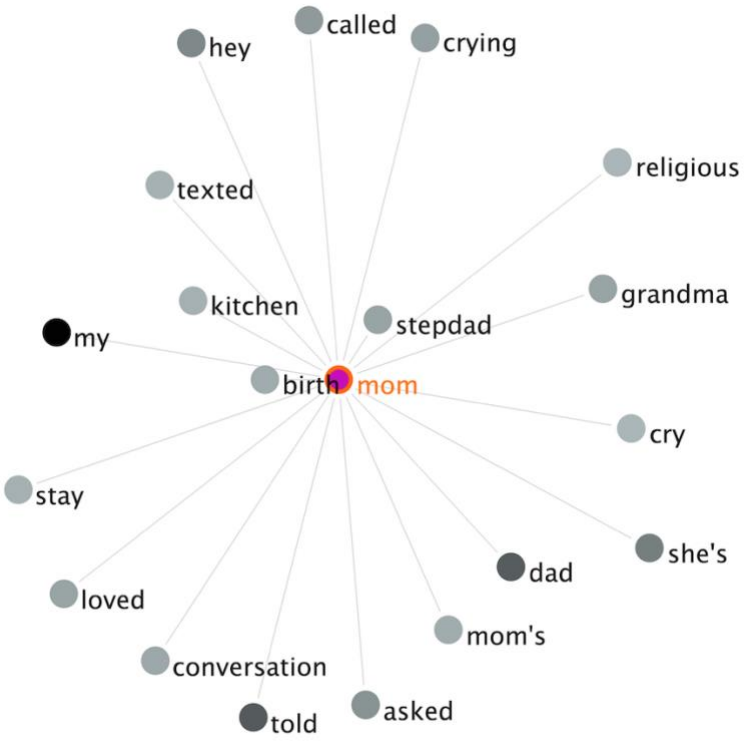

Figure 4. Collocates of Mom

Table 11 demonstrates that the five strongest collocates of the node **dad** proceed as follows: *loves* (5 co-occurrences), *brothers* (5), *mom* (70), *seen* (6), and *my* (524). Furthermore, similarly to the node **mom**, co-occurrence of the node **dad** and the collocate *my* is significantly higher than for other collocates indicated in the table: the frequency of *my* plus **dad** is 524, the next highest in the frequency measure collocation – **mom** plus **dad** – co-occurred 70 times. Thus, the occurrence difference between the former and the latter exceeds 7 times frequency.

Table 11

Collocates of Dad

Rank	Collocate	MI score	Frequency (collocation)	Frequency (corpus)
1	loves	6.53	5	44

Table 11 (continued)

2	brothers	6.21	5	55
3	mom	6.15	70	804
4	seen	5.99	6	77
5	my	5.93	524	7017
6	room	5.77	10	150
7	both	5.77	9	135
8	called	5.54	9	158
9	he's	5.52	24	427
10	dad	5.49	26	471
11	knows	5.41	5	96
12	told	5.23	41	890
13	later	5.08	7	169
14	sister	5.04	7	174
15	literally	5.00	7	178
16	accepting	4.90	5	137
17	found	4.80	7	205
18	said	4.77	29	867
19	work	4.73	6	184
20	yes	4.72	6	186

Figure 5 is representative in detecting “magnetic force” between the note and collocates that measures in MI. Shorter lines here link the strongest collocations. Visual exemplification in Figure 5 is consistent with the data in Table 11: the strongest collocations are **dad** plus *loves* (MI 6.53), *brothers* (6.21), *mom* (6.15), *seen* (5.99), and *my* (5.93). Furthermore, the darker shades of circles (for instance, for the collocates *my* and **mom**) imply the higher frequency rank of the collocation in the corpus. Table 10 confirms this: the frequency of the co-occurrence of *my* and **dad** is 524 and of *mom* and **dad** is 70. As opposed to, for example, considerably lighter shade of the circle *love* with the frequency being only 5 co-occurrences.

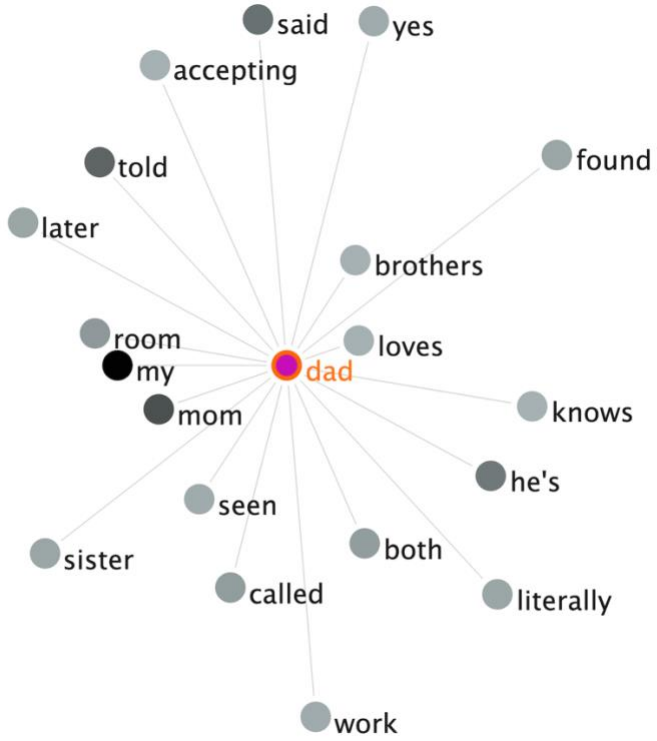

Figure 5. Collocates of Dad

Significantly high frequency of such collocations as **my mom** and **my dad** suggests that the vloggers attach particularly great importance to coming out to their parents. The purpose of this study does not imply analyzing the whole entirety of the collocations, thus concordance lines 26-37 embody some exemplary snapshots of the usage of **mom** and **dad** in the corpus. I start the analysis with the two most frequent collocations consisting of the nodes **mom** and **dad**: **my mom** and **my dad**.

Youtubers employed **my mom** and **my dad** to address a wide range of topics, which can be inferred from the prodigious frequency of the collocations. More specifically, men shared both negative and positive experiences of coming out to parents. Line 26 and 27 portray particularly negative reactions from mothers. The woman in line 26 was so upset with the news

of her son's homosexuality, so she performed an act of physical assault on her son for the first time since he was born: "It was the first time **my mom** had ever laid a hand on me, ever-ever hit me."

26. She's doing the laundry, and I said: "Hey! I'm gay!" And, basically, I actually got the living beat out of me. I used to have like model cars on my dresser… It was the first time **my mom** had ever laid a hand on me, ever-ever hit me. Um, I had to crawl up in the little ball, because she was smashing glass everywhere. So that was kind of painful to me because I then had bruises all over me.

Furthermore, the mother in the line 27 determined that homosexuality of her son is something that undermines all his life's accomplishments or even her ability of being a good mother: "She said that I was her biggest disappointment." The son consequently reaches the conclusion that she no longer loves him: "My mom, who I thought would love me and care about me no matter what, didn't."

27. It, kind of, sucked because **my mom,** who I thought would love me and care about me no matter what, didn't. And she said that I was her biggest disappointment.

Concordance line 28 appears to demonstrate that the vlogger was thinking that for his father, who had a conservative upbringing, the coming out process may not go smoothly: "**My dad,** I knew he'd be the hardest." The father suggested that his son's homosexuality is temporary: the father "said that he thinks it's just a phase." A provisional nature of homosexuality together with assumptions that there can be a choice whether or not to become gay are frequent topics within the coming out stories in the corpus. For instance, in line 29 we

see: “My mom got more anxious because, you know, she believed being gay was a choice and being gay was bad.”

28. I just knew that my family still loved me, my mom my sister… It was all great. So then came **my dad**. **My dad,** I knew he'd be the hardest because **my dad** born and raised on a farm in the country, just like an old time an old soul… He said that he thinks it's just a phase.
29. As I got older, [I] started, you know, becoming more and more feminine. **My mom** got more anxious because, you know, she believed being gay was a choice and being gay was bad.

More positive coming out to parents experiences can be observed in the concordance lines from 30 to 33. More specifically, lines 30 and 32 have to do with especially emotional reactions that took place during the process of coming out. For instance, one of the vloggers describes his coming out to mother as “the most emotional experience of my life” (Line 30). Another man simply could not articulate the confession about his homosexuality, but instead he put his head on his mother’s lap in seeking comfort. In line 31, the phrase “I felt so vulnerable…” suggests that regardless of one’s age, the load of stress attached to the coming out process may emotionally transform a man to a “vulnerable young child.”

30. For me, the most emotional experience of my life was coming out to **my mom.** And I don't know if that's because my life has been extraordinarily easy or if because coming out to your parents is…
31. When **my mom** asked what I needed to say, I couldn't find the words as she held my head in her lap, I felt so vulnerable like a young child.

Furthermore, there is an ongoing narrative occurring in the corpus that deals with the transition of parents' attitude towards homosexual sons. As we could observe, the initial reactions of some of the parents are on a spectrum of denial or disapproval. Nevertheless, it is not uncommon when parents change their outlooks with time. Line 32 demonstrates such a transition: "I've seen him [father] go from 'okay' to like 'my son's gay, he's gay, that's it.'" The vlogger asserts that his father invested a great deal of effort in trying to understand his son's sexuality. Another not uncommon narrative throughout the corpus is when parents knew or surmised that a son is gay but for some reason never raised the issue first. Line 33 exemplifies such a narrative.

32. I've seen him go from 'okay' /suspicious intonation/ to like 'my son's gay, he's gay, that's it.' And so, I… I love **my dad**. He's been such a great… He's given such a great effort and he's come so far, and it's just amazing to see that like he loves me.
33. I just never really get to it. I was just chickened out. And **my mom** came up and the first thing she said, she was like: "I've always known."

Mom and dad is another frequent collocation that is being employed by youtubers in a number of different themes. In general, I was not able to trace any specific roles attached to the collocation **mom and dad** that have not been already used for **mom** or **dad** separately. I would assume that relationship with parents in general as well as the need of approval from parents is a driving force behind such high frequencies of the collocations. As can be observed in line 34, there is a fear that the son's gay-related browser history might be revealed. Line 35 illustrates a rather unusual narrative of coming out first to parents and after that to somebody else. More

often in the corpus, gay teenagers and young men come out to friends or siblings before they do to their parents.

34. I was so ashamed… so ashamed! and I was very-very careful. I was like **mom and dad** could see this. So, I learned all of those tricks, you know, all the tricks of deleting history.
35. I've been thinking about coming out to someone other than my **mom and dad** because as far as it's been for the past couple of months, I've only been out to them.

Among the strongest collocates of the node **mom**, there are **birth** and **stepdad.** It can be explained by the existence of assorted configurations of parenting after parents get a divorce. Sometimes, sons stay with the biological father and his new wife (line 36) or with the biological mother and stepfather (line 37).

36. I lived in **my dad** and **stepmom**'s house but every other weekend I would go over to my **birth mom** and **stepdad**'s house.
37. "But you don't want to be a part of your family." And that was really the end of the conversation. My **stepdad** and **birth mom** and their whole fricking messed up Church.

God and Religious

In the previous subsections of the current chapter, I attempt to describe and discuss the nodes that were selected based on their prominence within the major keyword themes tables. However, I believe it is essential to pinpoint and examine not only the most frequently represented key-words but also the lexemes that occur less frequently or even transcend the

yielded lists. In this regard, the topic of religion is worth pointing out. This theme is constructed by only one single-word keyword (**god**) and just two phrases in the list of multi-word keywords (*religious family* and *conversion therapy*). At the same time, for example the word **god** has relatively high frequency (18th ranking), and as we just discovered, it singlehandedly creates the entire new theme (under the conditions of the current research design). This role that the node **god** plays in the corpus may account for the node's particular importance for some of the vloggers that made a decision to use it. In addition, I will analyze the most frequent item from the same theme of the multi-word keyword table – the node **religious** as in *religious family*.

Tables 12 and 13 each incorporate twenty collocates with the highest MI scores linking the collocates and the nodes **god** and **religious**, respectively.

Table 12

Collocates of God

Rank	Collocate	MI score	Frequency (collocation)	Frequency (corpus)
1	bless	9.78	7	9
2	prayed	8.82	6	15
3	created	8.34	6	21
4	testimony	7.94	5	23
5	wants	7.86	7	34
6	oh	7.77	135	698
7	child	7.73	9	48
8	happens	7.30	5	36
9	pray	7.14	5	40
10	believe	6.69	13	142
11	thank	6.19	11	170
12	relationship	5.92	10	187
13	god	5.73	16	340
14	has	5.52	10	247
15	wrong	5.14	5	160

Table 12 (continued)

16	take	5.11	8	262
17	made	5.00	9	319
18	literally	4.99	5	178
19	am	4.97	15	542
20	us	4.95	5	183

As can be observed in Table 12, the five strongest collocates of the node **god** are distributed as follows: *bless* (7 co-occurrences), *prayed* (6), *created* (6), *testimony* (5), and *wants* (7). It is particularly noteworthy that the co-occurrence of the node *god* and the collocate *oh* is considerably higher than this of others shown in the table. More specifically, the frequency of *god* plus *oh* is 135, whereas the next highest in the frequency measure collocation – *god* plus *am* – co-occurred 15 times. Thus, the frequency difference between the former and the latter is multiplied 9 times. Figure 6 provides the visual representation of the phenomenon: the collocate *oh* appears to have the darkest shade among other collocates.

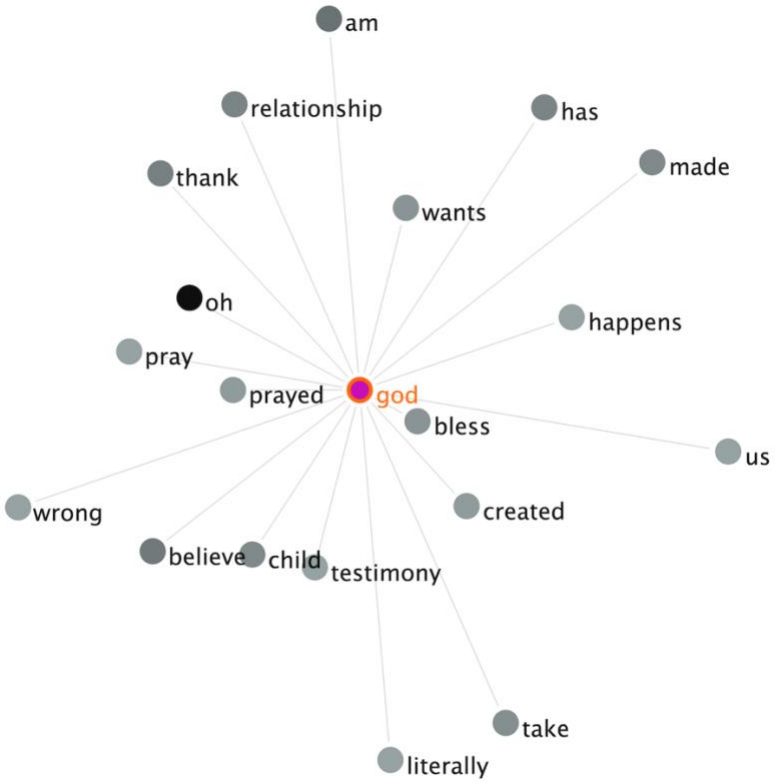

Figure 6. Collocates of God

The comparing in magnitudes of the force that pulls the collocates to the node is visually exemplified in Figure 6. Precisely speaking, the five shortest lines that correspond with the five strongest MI's link the node *god* and the collocates *bless*, *prayed*, *created*, *testimony*, and *wants.* Their MI's rank from 9.78 for *bless* to 7.86 for *wants*.

Table 13

Collocates of Religious

Rank	Collocate	MI score	Frequency (collocation)	Frequency (corpus)
1	super	8.10	7	156
2	very	7.41	23	825
3	she's	6.80	6	328
4	family	6.76	10	564

Table 13 (continued)

5	they're	6.36	5	371
6	from	5.29	5	778
7	mom	5.25	5	804
8	not	4.90	11	2254
9	are	4.89	6	1239
10	really	4.88	11	2277
11	because	4.60	9	2265
12	up	4.55	5	1305
13	people	4.51	7	1872
14	but	4.50	12	3237
15	a	4.15	19	6534
16	with	4.01	6	2279
17	is	3.95	6	2373
18	about	3.94	5	1996
19	my	3.61	14	7017
20	that	3.50	15	8093

Table 13 and Figure 7 illustrate the relationship between the node **religious** and its collocates. We can observe that on the top of both MI and Frequency (within collocation) scores *very* occupies leading positions: 7.41 and 23 occurrences, respectively. Thus, I decided to select the combination *very* plus **religious** for the concordance analysis. Furthermore, *super* plus **religious** will be analyzed as the strongest collocation of the node, with the MI scoring as high as 8.10. The collocate *family*, due to its significantly prominent bonding with the node under consideration, will be analyzed further.

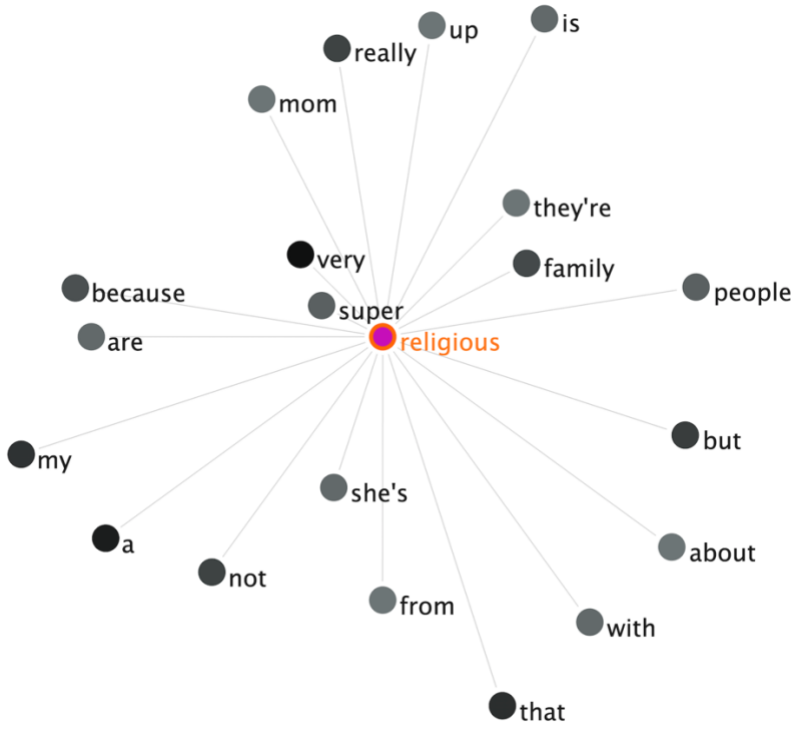

Figure 7. Collocates of Religious

The collocate *oh* has the largest number of co-occurrences together with the nod *god,* mostly in such phrases as **oh god** or **oh my god** (see line 38 as an example for the latter). The phrase **oh my god** by itself yields 109 results in the concordance list (out of the total 135 co-occurrences of *oh* plus **god**). At the same time, neither **oh god** nor **oh my god** appears in the multi-word key-word list. In fact, **oh my god** is not in the list of the first one thousand multi-word key-words (I checked the extended version of the list specifically for the collocation). This means that even though the phrase is used fairly often in the corpus, its frequency measured against the size of the corpus does not exceed its overall popularity in English Web corpus 2015 that contains fifteen billion words.

38. I hear footsteps downstairs: she's coming up the stairs. **Oh my god!** What am I gonna do?

Furthermore, the node **god** may be used to justify the very existence of a gay person. There is an ongoing debate between the people who believe that being gay is a human predetermined condition (in other words, people are born gay) and those who believe that sexuality is a choice. According to Gessen (2019), the fact that sexuality cannot be changed, and gay people are simply powerless to do anything about it, is a premise underlying the gay rights movement. Some religious gay vloggers support this point view. For instance, in the line 39, we observe that the vlogger is supporting the predetermined nature of homosexuality by saying "**God** doesn't make mistakes… I firmly believe **god** made me this way."

39. **God** doesn't make mistakes… I firmly believe **god** made me this way, and this is my story.

The same notion is expressed by a particularly strong collocation **god created**. For instance, in the line 40, the man asserts that "if you're gay, **god created** you gay." Further, in the line 41, the vlogger takes a step forward in the discussion about the reasons that predetermine homosexuality and reflects that being gay is "about me being brave and bold, to fully be the man that **god created** me to be." First of all, the man endorses the viewpoint that he was born gay, not that he chose to be that way. And secondly, he introduces the notion that being gay is not a pleasant or desirable task ("this is not about my glory") but rather an ongoing fight or struggle that requires him to be brave and bold.

40. I don't believe god would heal such things, you know, because I think, if you're gay, **god created** you gay. You know, that's just the way it is.

41. This is not about my glory. It's about me being brave and bold, to fully be the man that **god created** me to be.

God bless is another strong collocation with the MI scoring as high as 9.78. It is worth pointing out that out of seven co-occurrences of the words in the corpus, three are dedicated specifically to the audience. Vloggers decided to finish the videos with the words of blessing: "I pray that someone has been touched by this video. **God bless** you" (line 42) and "Have a wonderful night. May **god bless** you."

42. Well that's all the time I have for today. So, I pray that this blesses you, I pray that someone has been touched by this video. **God bless** you.

43. So, yeah, thank you all so much! I will see you all later. Have a wonderful night. May **god bless** you. And, as always, I will see you in the next video.

Throughout the concordances that have to do with the strongest collocates of the node **religious** – **super**, **very**, and **family** – I discovered a plethora of discourses colored in negative undertones. To be more specific, lines 44-46 portray instances when vloggers had to find a way to deal with **very religious** or **super religious** members of their families. In line 44, a man creates a causal link bonding the fact that his grandparents are very religious and, therefore, are homophobic: it was challenging to come out to them because the grandparents are "**very religious** people, and they don't really accept the fact that there are homosexuals around."

44. Coming out to my grandparents on my dad's side – that was harder. That was much harder to do because they are **very religious**, they're **very religious** people, and they don't really accept the fact that there are homosexuals around.

Some gay men start their coming out journey with coming out as bisexuals. At times, even this sort of "hedging" does not work out the way the men desire it to. For instance, in line 45 the vlogger compares opening up to his parents with the form of a physical assault: "me coming out [to my parents]… – it's like a slap in both of their faces."

45. I still had the mentality where I was like: this cannot happen for me. If you watched my last video, my mom is **very religious**, and my dad is very masculine. So, me coming out and saying: "Hey! I like both sexes!" – it's like a slap in both of their faces.
46. I kind of was like attracted to men but never really pursued it because I grew up in a **very religious family**, and they wasn't very accepting, you know.

In line 47, the vlogger makes a decision to include such personal trait as **super religious** in the list of characteristics which may become detrimental to one's safety. More specifically, he states: "If your family is not accepting gay people and, you know, maybe **super religious**, or, for whatever reason, your personal safety would be threatened by coming out."

47. If your family is not accepting gay people and, you know, maybe **super religious**, or, for whatever reason, your personal safety would be threatened by coming out, honey, just hold onto this, wait until you're eighteen move out and then start over…

In Summary

Lovelock (2017) asserts that *YouTube* Coming Out stories enable queer population to articulate what their day to day experience of living in a straight world feels like. My research

uncovers a set of specific themes that are instrumental in describing such experiences. Through the analysis of the thematic groups of my corpus, one can tackle such questions as

- When does one find out he is gay?
- What are people in the closet afraid of when it comes to coming out publicly?
- What role does a family play in the process of accepting one's sexuality?
- How to deal with omnipresent homophobia?
- Etc.

The major themes that emerged from the keyword items coding are the following: *Family, Education, Relationship, Social Media, Vlogging, General Gay-Related Items, Sexuality, Coming Out, Profanity, Homophobia, and Religion*. My research design does not imply examination of each keyword's collocations and concordances, yet even the most frequently occurring items, that I had decided to work with, provided abundance of information on the subjects that were chosen worth mentioning in the videos.

More specifically, it turns out YouTubers tend to answer "When does one find out he is gay?" question by linking sexuality awareness stages to their school or college years. A freshman year of college, for instance, is not just a beginning of a new educational phase but also the time of the brand-new freedom and fresh experiences. Having emancipated and surrounded themselves with entirely new social circles, a lot of young men finally decide to come out. The collocation **freshman year** occurred 72 times, which makes it the most frequent collocation of the *Education* theme. However, the topics related to mentioning freshman year of high school or college are not confined solely to public coming out. Vloggers employ the collocation to pinpoint various events of their lives from bad grades to first romantic relationships. The

collocations containing the node **grade** are dealing mostly with elementary and middle school years. Consequently, the node can be found within different contexts that are not related to actual coming out but mostly have to do with realization of one's homosexuality.

Collocation **my mom** and **my dad** occur in the corpus 852 and 524 times, respectively, which makes them the most popular collocations within the key-words selection. Wuest (2014) states that gay kids had to create safe spaces to share information and experiences in the form of YouTube Coming Out stories, among other things, because a great deal of homophobic parents still believe that they can prevent their gay sons and daughters form being gay. As a result, children, who quickly realize it is impossible to simply "pray the gay away" no matter what, come to realization that there must be something terribly wrong with them. I my study, I discovered that vloggers have been particularly outspoken about relationships with their parents. The contexts of using **my mom** and **my dad** address a wide range of topics, namely, physical and verbal assaults as a result of coming out; unconditional parental love; change in opinion concerning homosexuality; total support; etc.

It is worth pointing out that the narrators uttered **my mom** more than 1.5 times as much as they did **my dad.** There can be a few explanations behind the discrepancy. First, it is possible that gay sons simply feel closer to their mothers than fathers. On the other hand, such concordance lines as "… **my dad,** I knew he'd be the hardest" might suggest that mothers are more likely to understand, support, or forgive, as opposed to fathers.

I can bring more possible reasons explaining the higher frequency of the collocation **my mom**; however, what I believe is particularly important is to narrow down the question "why do we want to know those reasons?" Through the analysis of data, I discovered that gay vloggers

tend to frequently mention lexical items that have to do with the topics of family, education, relationship, homophobia, religion, and so forth. On the other hand, YouTube Coming Out stories have been found particularly important and popular [Lovelock (2017), Pullen (2010), Losh and Alexander (2010), Wuest (2014)] among LGBTQ audience. In this regard, various researchers from different fields of study may consider conducting more in-depth investigation concerning why exactly the vloggers speak more about their mothers or why the topic of religion is in the top 12?

More distinctively, since more traditional social institution – namely, schools, churches, and families – fail to facilitate the level of support that is so desperately needed within the gay youth community (Wuest, 2014), social workers and psychologists can take a closer look at the YouTube stories. There must be something in those twelve (even more items might be added dependent on the particular research design) themes that helps and inspires gay audience to come out of the closet and film and post more videos. These themes can be addressed during round tables and counseling meetings with troubled and confused young queer individuals.

Craig and McInroy (2014) conclude that new media (websites, web-based TV, web-based news, social media, social networking, and video sharing) create critical opportunities for young LGBTQ population to pinpoint and reflect on their identities. Indeed, it is not always the case when queer teenagers easily come to realization about their homosexuality. In other words, growing up in a heteronormative reality is not conducive for a steady development of a queer identity. The bottom line is, according to Craig and McInroy (2014), various online activities, including YouTube sharing of Coming Out stories, facilitate processes that eventually help to come out publicly. Vloggers choose assorted ways to reach the point of leaving the closet. Some

of them spend hours of binge watching how other people manage to come to terms with their sexuality, others create fake social media accounts and join queer communities anonymously. In any case, vloggers create a network of like-minded people that are struggling with exploring gay identities. From this standpoint, corpus linguistic analysis of digital coming out narratives can contribute to the fields concerned with learning how such identities are being developed.

The question of discovering and developing queer identities has been tackled from various perspectives. One that has to do with corpus linguistics was illustrated by Baker (2003). He conducted corpus-driven analysis of personal advertisements from *Gay Times* (formerly *Gay News*) magazine published from 1973 until 2000. The value of personal adverts as important indicators that can be used to illustrate the development of homosexual identities lies within the assumption that such posts "are often expressions of idealistic desire" (p. 258). One of the major building blocks in the construction of identities, as demonstrated by Baker (2003), is fear of being perceived feminine.

The mode **feminine** turned out to be one of the key-words in this study. However, before making any comparisons, I would like to mention two important caveats. First of all, Baker's study deals with the written form of English used in a British magazine in the last three decades of the 20th century; whereas, my study is concerned with spoken American English of *YouTube* blogs posted through the 2000's and 2010's. Secondly, Baker (2003) did not make any distinctions between the adverts of gay men authorship and those written by other cis- or transgender individuals. My corpus, on the other hand, was formed off the utterances produced by gay men exclusively. With this being said, a more in-depth look at the node **feminine** through its concordances reveals that the narrators express negative connotations while producing the

node. More specifically, one of the examples that is charged with the same mood as many other lines is “As I got older, [I] started, you know, becoming more and more **feminine**. My mom got more anxious because, you know, she believed being gay was a choice and being gay was bad.”

It is important to specify that Coming Out videos aim to tell a story of obstacles on the path to finally leaving the closet. In other words, attitudes towards being feminine have been expressed through the lens of either heterosexual majority (as illustrated by the concordance line in the previous paragraph) or gay men while they are still in the closet. With the numerous onsets and rapid development of such shows as RuPaul's Drag Race (Visage & Polly, 2009–) and the emergence of multiple male YouTube and Instagram make-up artists and beauty bloggers, that have millions of followers from all over the globe, the attitudes towards effeminate men are not that strongly negative anymore. However, for gay men who are still in the closet, even suspicion, let alone accusation, of being not manly enough can have severe undesirable consequences.

Another substantial gay associated stigma mentioned by Baker (2003) is HIV/AIDS. In the 1980’s, gay men were the primary suspects of spreading HIV and subsequently causing the pandemic of AIDS. As a result, negative attitudes towards homosexuality was increasing through the 1980’s and early 1990’s. Apparently, HIV/AIDS rhetoric played an important role in gay ads narratives at that time frame and beyond and become one of the biggest fears among homosexual men. In my study, the node **HIV** occurs only twice, whereas **AIDS** – 5 times, 3 of which were used to provide with historical references. From these findings, it is safe to say that for the vloggers from my selection, HIV/AIDS issue is no longer a stigma or substantial threat associated with being gay.

When it comes to digital media, the timeline of coming out itself comes about in a different manner compared to older days. Before invention of *YouTube, Facebook, Snapchat,* etc., revealing person's sexuality or sexual orientation to others used to be attributed with the ultimate stage and/or goal of coming out process (Cover & Prosser, 2013). According to Lovelock (2017), however, YouTubers dedicate little time to coming out as such. As a rule, the audience presupposes that the vlogger who posted a Coming Out story must be part of LGBTQ. What the vloggers actually mean to talk about is the entirety of the stages of the coming out journey: before realizing, realizing, denial, struggle, acceptance, and so forth. My findings are consistent with those of Lovelock (2017). The narratives that I discovered through examination of assorted concordance lines, that deal with the matters of family, sexuality, education, religion, and others, suggest that the YouTubers are generally concerned with offering a "therapeutic" session rather than a story simply culminating with revealing one's sexuality, as if to say, "Yes, I am gay and likely you are, as well. Here is my story. I have been through a lot. Accepting yourself, let alone coming out to others, might seem hard or even impossible task right now. But it will get better because you are not alone."

Within more in-depth analysis of coming out narratives, Cover and Prosser (2013) pinpoint a theme of major significance. The authors emphasize the recurrence of such utterances as 'I remember', and 'when I was three' or 'when I was 13'. The point is that even in industrialized countries with highly developed evidence-based healthcare systems, general public has been split into two teams: "homosexuality is innate" and "it is a choice." And by making references to events and experiences of childhood, YouTubers, often unconsciously and implicitly, provide evidence in favor of "I was born gay" point of view as opposed to "I decided

to become one." In the current study, the narratives within the *Education* theme enable to scrutinize the development of realization of one's sexuality through the numerous reoccurring allusions and references to specific grades and college years. In this regard, some additional research is required since the earliest school-related age is five or six years old. It seems possible to search for an additional theme related to early childhood to complement the *Education* theme in order to create an age/school-years related summary of the formation of gay identity.

And finally, from the Queer Theory point of view, coming out videos play particularly important role in changing general public's attitudes towards LGBTQ community. Queer Theory seeks to be perceived as a lens or tool to question and review the contemporary oftentimes rigid ideas of social structures and taxonomies when it comes to individuals that do not comply with heteronormative status quo. One of the major tools to bring social changes about is to assert that the matter exists. Madden (2014) said "One of the most important things we can do to change our culture is to tell our stories... Telling stories matters because when I listen to your story, I not only feel with you and for you, I have to make decisions about how to treat you." By sharing their coming out stories, LGBTQ people not only articulate various life experiences, but also alter the fabric of society by creating and augmenting visibility of fellow queer individuals thus making heteronormative majority recognize that, regardless of sexual orientation differences, a great deal of the topics discussed in the videos (*Family, Education, Relationship, Social Media, Vlogging, Profanity, Religion, etc.*) are not strictly unique to LGBTQ. Straight people can listen to a coming out story and make a decision about how to treat gay people on the basis of similarities that might not seem obvious if queer people fail to assert themselves.

Chapter VI: Limitations and Further Research

When I was developing the research design for the current study, in order to answer to the research questions, among other things, I intended to conduct a diachronic analysis of the yielded frequencies and keywords. It turned out that the initial stage of the research process – forming the corpus – took much longer than was originally enshrined. Nonetheless, carrying out the diachronic analysis of the data might help to highlight assorted substantial linguistic trends. For instance, comparing corpora, that composed tailored to the dates when the coming out stories were posted (a pool of older videos versus a pool of the most recent ones), may help to pinpoint lexical items that undergo fluctuations in the frequency of use in time. More specifically, Baker (2003) employed diachronic analysis to reveal how gay men used language to construct identities and how such identities have altered over time.

Tracing the process of constructing identities does not necessarily have to be carried out diachronically. Another powerful tool that can be utilized for conducting a sociolinguistic study is critical discourse analysis (CDA). CDA views language use as a manifestation of social practice. Researchers involved in academic endeavors related to CDA attempt to determine the processes behind societal power interactions that are generated and reinforced through the means of language use. The nature of the research questions identified for the current study does not involve conducting CDA, yet the framework of my findings resembles that of Brindle's (2016). Brindle begins his study with corpus linguistics, identifying frequencies, key-words, collocations, and concordances; nevertheless, he does not stop at the stage of the concordance lines discussion but proceeds with using the findings for CDA. With an additional set of research questions, my study also can be further extended in order to underline issues of societal structural

inequities that originate within the interactions between the LGBTQ+ community and the heteronormative majority.

Another substantial dimension for the further analysis can be enabled by creating collocational networks (CN's) based on the keyness of frequency measures. Some words – 'nuclear nodes' – attract higher number of collocates than others thus launching a network of collocation. The current study is designed to yield predominantly descriptive deductions regarding the corpus. At the same time, the keywords that have been identified through the analysis of data can be further rearranged into a CN. Such networks are very visual thus easy to comprehend. In addition, depending on the particular research aims, a CN can be created on the fly. Let us take a look at the relatively simplistic network comprised of the most frequent items from each semantically significant thematic categories within the framework of single-word key-words. MI is set at 5.0. There are two items – **god** and **bitch** – that did not yield collocates with other keywords, thus, they were excluded from the chart. The key-words (nodes) are presented in the rectangle segments; the words that have been yielded by the software as links, considering MI 5.0, are presented in ovals.

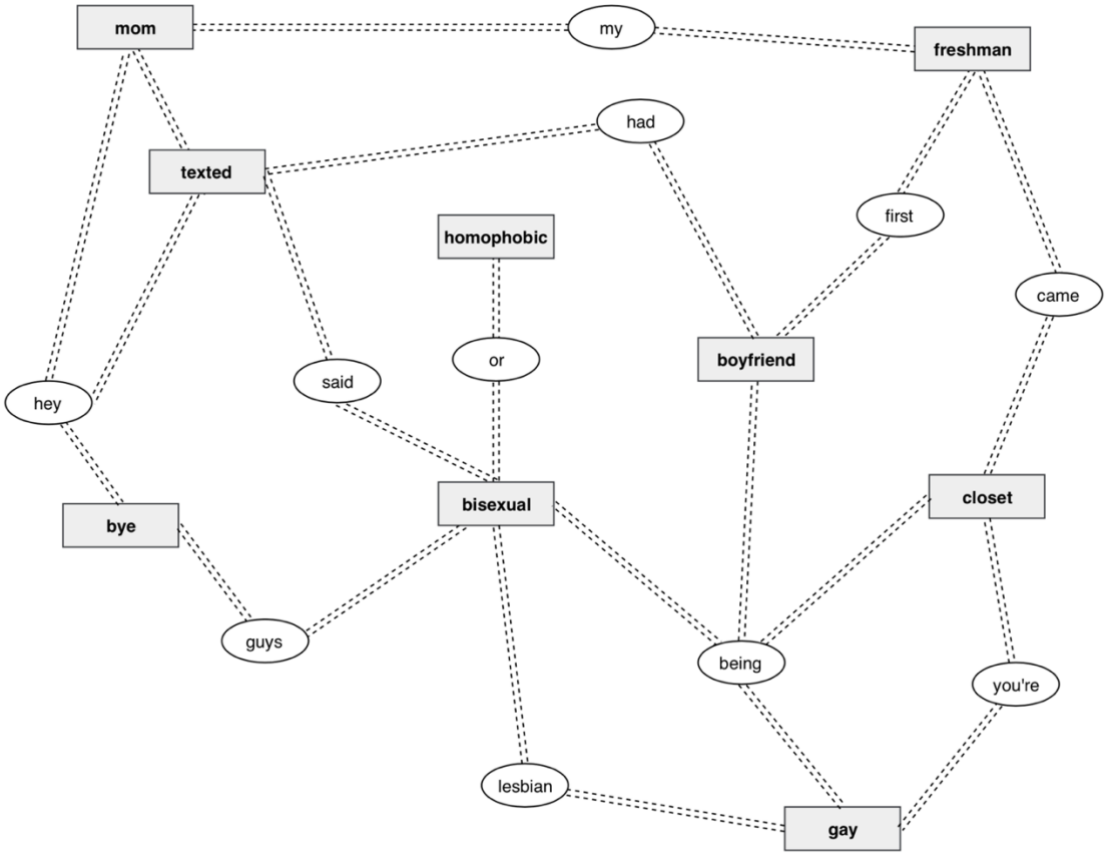

Figure 8. Collocational Network

With the unaided eye, in the Figure 8, we can observe that for the selected sample, *bisexual* and *texted* are the strongest nuclear nodes, meaning that they attract the largest number of collocates among other popular keywords. Note, that **bisexual** uses two link words to establish a collocation with the node **gay**. According to Brezina et al. (2015), CN is a convenient analytical tool that can be employed in assorted domains "of linguistic and social research such as discourse studies, psycholinguistics, historical linguistics, second language acquisition, semantics and pragmatics, lexicogrammar, and lexicology" (p. 165). In this regard, I see how CN's can find potential implementation as a supplement for more in-depth CDA of the corpus.

Chapter VII: Conclusion

My research questions aim to examine distribution of lexical words of the corpus and emerging major thematic groups from the yielded keywords. After the preliminary work on creating the corpus, I pursued the questions with creating tables that incorporate frequencies of lexical words and lemmas found in the corpus. Such tables are called frequency lists. Analysis of frequencies allows for identifying linguistic traits that can be observed repeatedly in corpora thus revealing patterns of language use. In my corpus, the most frequent word as well as lemma is 'I' which occurs 26,912 times. One of the possible explanations of such a popularity is that the corpus producers were concerned with sharing predominantly deeply personal narratives.

Furthermore, the vloggers were not shy to speak explicitly about their sexuality using the lemma 'GAY' 2112 times. There are some other patterns that have been discovered through the analysis of the frequency lists. For instance, the lemma 'GUY' has been employed more than twice as much as lemma 'GIRL': 1414 and 587 times, respectively. Considering that the corpus is composed by the men that are both romantically and sexually attracted to other men, such a discrepancy is not a surprise. Moreover, the lemma 'MOM' (863 occurrences) is presented in the sample considerably high – the twentieth ranking; the lemma 'DAD', on the other hand, is completely absent.

Lack of words in the frequency list can be used to derive implicit conclusions about the patterns of language use in the corpus. For instance, it is safe to say that the absence of the lemma 'DAD' in the selections speaks for the fact that mothers may play more important roles in their son's coming out process. In the same fashion, it is noteworthy that the lemma 'WOMAN' is not presented in the list at all. Furthermore, among the lexical items that express feelings, *love*

and ‘LOVE’ are not just scoring high (twentieth ranking for the words and twenty first – for the lemmas) but the only representatives of their kind in both frequency lists. The lack of the lemma ‘HATE’ implicitly proves that such a feeling does not bear a high value within the context of the coming out narratives. Nevertheless, one of the major limitations of the frequency analysis is that it does not allow for the in-depth examination of the topics that have been covered, for instance, using the lemma ‘GAY’ so many times.

Another corpus linguistics instrument showcases the distribution of the lexical items, however, from a different angle. Here, I am referring to keyword analysis, which is based on matching two corpora. For this study, I compared my corpus against the reference corpus, English Web corpus 2015 (enTenTen15). In this way, keywords are the lexical items whose frequency is significantly high in comparison with a reference corpus. *Sketch Engine* enables extracting two types of keywords: single-word and multi-word ones.

For the analysis of keywords, I employed general qualitative coding. The coding allows for identifying major thematic categories off the keywords lists. I discovered that gay men view as particularly important for the purposes of sharing with the audience the following twelve themes: *Family, Education, Relationship, Social Media, Vlogging, General Gay-Related Items, Sexuality, Coming Out, Profanity, Homophobia, Religion,* and *Other*. The most prominent groups, meaning those containing the highest numbers of keyword items, turn out to be *Family* for the single-word keywords and *Education* for the multi-word ones. Whereas, one of the smallest category is *Religion*.

Frequencies, keywords, and thematic groups enable a broader outlook on the corpus. That is to say, after rearranging texts with the aid of *Sketch Engine* and *#LancsBox*, it is possible to

conclude what topics have been selected by *YouTube* vloggers for the particularly sensitive in nature videos. However, what frequencies and keywords provide with is not simply a better understanding of linguistic patterns within the corpus; the instruments also give the base and justification for indicating more in-depth categories of corpus linguistics, namely collocations. Collocations, in turn, are being further illustrated and analyzed by the means of concordances.

Collocations are two or more words that co-occur more frequently that would be expected by chance. For the purposes of sociolinguistic studies, *characteristic* collocations have been frequently employed. This type of collocations helps to reveal semantic associations and connotations of co-occurring lexical items, thus pinpointing more specific characteristics of corpora than keywords and frequencies, for instance. To be more specific, we already identified one of the most frequent lexical words of the corpus: 'gay.' The word scores high in both the frequency and the keyword lists. Finding collocates of this node can take us one step further in understanding the associations in which the node has been employed.

One of the most frequent collocation formed with the node **gay** is 'being gay'. The collocation occurs in the corpus 271 times. We can assume that since the gerund 'being' is generally used to express experience or condition, it is particularly important for the vloggers to talk about what it means, feels, looks like, etc. to be gay. Now, even though we have learned more about the corpus, the significance of the collocation should not be overemphasized, since the context of the collocation is still missing. And here is the stage of the analysis where concordances come to the fore.

Concordance is a line that contains the node (word or phrase) positioned in the middle surrounded by the context of several words to the left and to the right of the node. Detailed

examination of a few concordance lines that contain the collocation ‘being gay’ as a node, reveals a plethora of semantic information about the context within which the phrase is situated. More specifically, the collocation has been used in the corpus to express:

- denying a person’s queerness and even hatred towards himself;
- desire to change the sexuality with help of a prayer or miracle;
- neutral attitude towards one’s homosexuality;
- positive effects of coming out;
- etc.

Based on the narratives above, I may conclude that collocation ‘being gay’ is an important instrument adopted by gay vloggers to communicate their attitudes about homosexuality, evaluate concerns and struggles, share joy and relief, etc. Further, in the same fashion as for the node ‘gay’, I analyzed collocates and concordances consisting the nodes **year** and **grade**; **mom** and **dad**; and **god** and **religious**.

In this chapter, I attempt to briefly guide the audience through the process of corpus linguistics that I employed in the current study. Four powerful instruments of the analysis – frequencies, keywords, collocations, and concordances – allowed for moving through the journey from having raw data to exploring patterns of language use, major thematic groups, word associations, and significant contexts built by the lexical items under consideration. The results yielded from the analysis can help to draw conclusions regarding the phenomenon of coming out narratives through lenses of assorted fields (Critical Theory, Psychology, Education, Popular Culture, etc.) and social strata (LGBTQ+ spectrum and heteronormative majority).

References

Baker, P. (2003). No effeminates please: A corpus-based analysis of masculinity via personal adverts in Gay News/Times 1973–2000. *The Sociological Review*, *51*(1_suppl), 243-260.

Baker, P. (2004). 'Unnatural Acts': Discourses of homosexuality within the House of Lords debates on gay male law reform. *Journal of sociolinguistics*, *8*(1), 88-106.

Baker, P. (2016). The shapes of collocation. *International Journal of Corpus Linguistics*, *21*(2), 139-164.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 31-40.

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, *20*(2), 139-173.

Brindle, A. (2016). *The Language of Hate: A Corpus Linguistic Analysis of White Supremacist Language*. Routledge.

Bucholtz, M. (2004). Introduction in *Language and woman's place: Text and commentaries* (Vol. 3) by Lakoff, R., & Lakoff, R. T.

Cameron, D. (Ed.). (1998). *The feminist critique of language: A reader*. Psychology Press.

Caskey, D. M. (2011). *Speak like a wo(man): A corpus linguistic and discourse analysis of gendered speech* (Doctoral dissertation, Western Carolina University).

Charmaz, K. (2008). Grounded theory as an emergent method. *Handbook of emergent methods*, *155*, 172.

Cheng, W. (2011). *Exploring corpus linguistics: Language in action*. Routledge.

Cover, R., & Prosser, R. (2013). Memorial accounts: queer young men, identity and contemporary coming out narratives online. *Australian Feminist Studies*, *28*(75), 81-94.

Craig, S. L., & McInroy, L. (2014). You can form a part of yourself online: The influence of new media on identity development and coming out for LGBTQ youth. *Journal of Gay & Lesbian Mental Health*, *18*(1), 95-109.

Gaudio, R. P. (2004). The way we wish we were: Sexuality and class in Language and Woman's Place. *Language and woman's place: Text and commentaries*, 283-288.

Gessen, M. (2019, April 8). Pete Buttigieg Claims His Right to Run for President – and Defends His Right to Exist. *The New Yorker.* Retrieved from https://www.newyorker.com/news/our-columnists/pete-buttigieg-claims-his-right-to-run-for-presidentand-defends-his-right-to-exist

Gray, J. (1992). *Men are from Mars, women are from Venus: a practical guide for improving communication and getting what you want in your relationships.* NY: HarperCollins.

Johnson, F. L. (1983). Political and pedagogical implications of attitudes towards women's language. *Communication quarterly*, *31*(2), 133-138.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, *1*(1), 7-36.

Kulick, D. (2000). Gay and lesbian language. *Annual review of anthropology*, *29*(1), 243-285.

Lakoff, R. (1973). Language and Woman's Place. *Language in Society*, *2*(1), 45–80. Retrieved from http://www.jstor.org/stable/4166707

Lakoff, R. & Lakoff, R. T. (2004). *Language and woman's place: Text and commentaries* (Vol. 3). Oxford University Press, USA.

Lassner, A., Glavin, E., DeGeneres, E., Paratore, J., & Connelly, M. (Executive Producers). (2003–). *The Ellen DeGeneres Show* [Television show]. Los Angeles: Warner Bros. Television.

Leap, W. L. (2015). Queer Linguistics as Critical Discourse Analysis. In *The Handbook of Discourse Analysis* (661-680). Blackwell.

Losh, E. & Alexander, J. (2010). "A YouTube of One's Own?": "Coming Out" Videos as Rhetorical Action. In *LGBT Identity and Online New Media* (pp. 51-64). Routledge.

Lovelock, M. (2017). 'My coming out story': Lesbian, gay and bisexual youth identities on YouTube. *International Journal of Cultural Studies*, 1367877917720237.

Lüdeling, A., & Kytö, M. (Eds.). (2008). *Corpus linguistics* (Vol. 1). Walter de Gruyter.

Madden, E. (2014, October). *Thinking out loud: on dangerous books, difficult stories, different lives.* Keynote presentation for South Carolina Library Association, Columbia, SC.

Martin, D. (2002). *Allen Read, 96, the 'O.K.' Expert, Is Dead*. Retrieved from https://www.nytimes.com/2002/10/18/nyregion/allen-read-96-the-ok-expert-is-dead.html.

Motschenbacher, H., & Stegu, M. (2013). Queer Linguistic approaches to discourse. *Discourse & Society*, *24*(5), 519–535.

O'Keeffe, A., & McCarthy, M. (Eds.). (2010). *The Routledge handbook of corpus linguistics*. Routledge.

Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc* (Vol. 10, No. 2010, pp. 1320-1326).

Piantato, G. (2016). How has queer theory influenced the ways we think about gender? *Working Paper of Public Health*, *5*(1).

Pullen, C. (2010). The murder of Lawrence King and LGBT online stimulations of narrative copresence. In *LGBT identity and online new media* (pp. 17-36). Routledge.

Radosh, D. (2004). *Why Know?* Retrieved from *https://www.newyorker.com/magazine/2004/12/06/why-know*

Rodgers, B. (1972). *The queens' vernacular: A gay lexicon*. Straight Arrow Books.

Tannen, D. (1990). *You Just Don't Understand.* New York: Ballantine Books.

Taylor, C. (2008). What is corpus linguistics? What the data says. *ICAME journal*, *32*, 179-200.

Tech Consultant (2017, January 3). *The Queens Vernacular A Gay Lexicon Dr. Judith Reisman.* [Vide file]. Retrieved from https://www.youtube.com/watch?v=euU7UnIKahg

Upton, T. A., & Connor, U. (2001). Using computerized corpus analysis to investigate the textlinguistic discourse moves of a genre. *English for Specific Purposes*, *20*(4), 313-329.

Visage, M., & Polly, J. (Producers). (2009–). *RuPaul's Drag Race* [Television show]. NYC: H1

Wolfram, W. & Schilling-Estes, N., eds. (2006). *American English: dialects and variation* (2nd ed.). Malden, Massachusetts: Blackwell Pub.

Wuest, B. (2014). Stories like mine: Coming out videos and queer identities on YouTube. In *Queer youth and media cultures* (pp. 19-33). Palgrave Macmillan, London.

Zimman, L. (2009). 'The other kind of coming out': Transgender people and the coming out narrative genre. *Gender & Language*, *3*(1).